SESSION

INFORMATION AND KNOWLEDGE EXTRACTION, REPRESENTATION AND SHARING

Chair(s)

TBA

Applying a Semantic & Syntactic Comparisons Based Automatic Model Transformation Methodology to Serve Information Sharing

Tiexin WANG, Sebastien TRUPTIL, and Frederick BENABEN
Centre Genie Industriel, University de Toulouse - Mines Albi, Albi, France

Abstract - *Information sharing, as an aspect of information and knowledge engineering, attracts more and more attention from researchers and practitioners. Since a large amount of cross-domain collaborations are appearing, exchanging and sharing information and knowledge among various domains are inevitable. However, due to the vast quantity and heterogeneous structures of information, it becomes impossible to maintain and share cross-domain information relying mainly on manual effort. To enhance the efficiency of sharing information, this paper presents an automatic model transformation methodology. Comparing to traditional model transformation methodologies, this methodology shields the general weaknesses: low reusability, contain repetitive tasks and involve huge manual effort, etc., by combining semantic and syntactic checking measurements into a refined transformation process. Moreover, the semantic and syntactic checking measurements are supported by software tool; in this way, manual effort is replaced from the information sharing/exchanging process.*

Keywords: Information sharing, automatic model transformation, semantic and syntactic checking

1 Introduction

Data, as the basis of information and knowledge, are "symbols that represent properties of objects, events and their environments" [1]. Data are products of observation"; furthermore, "information is contained in descriptions, answers to questions that begin with such words as who, what, where, when and how many". Information systems generate, store, retrieve and process data" and "information is referred from data". Fig.1 shows the relationships among data, information, knowledge and wisdom.

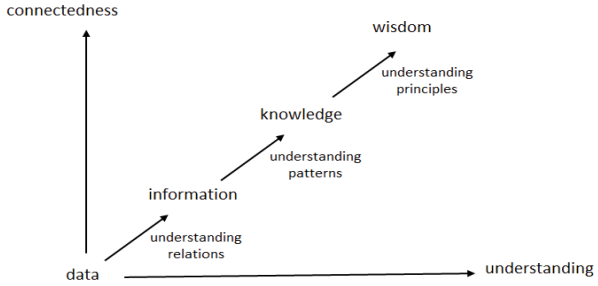


Fig. 1 Different presentation layers from data to wisdom [1]

Data is raw; it can exist in any form without significance. Information is generated by adding meaning (relational connection) on data, for example: data stored in a relational database could be regarded as information. Knowledge, which exists on a higher understanding level than information, is "a deterministic process" [1].

Based on the context of information sharing, Fig.2 shows the relationships among "nature languages", "data & information" and "specific domains".

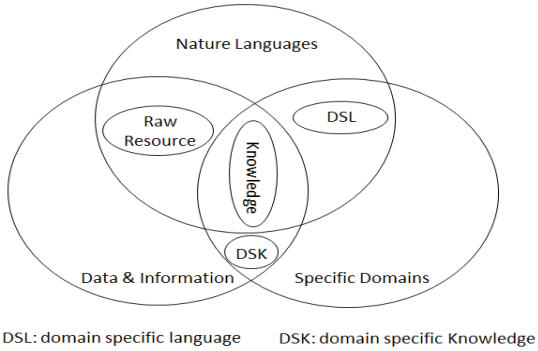


Fig. 2 nature languages, data & information and specific domains

Data and information are presented mainly by nature languages (also with the help of mathematical symbols, diagrams, etc.). Information applied on specific domain, could be regarded as knowledge; while this kind of knowledge may be regarded as information (or data) by other domains.

Due to the Internet, vast amount of heterogeneous data and information appear and they are provided by different sources. Three typical sources are: Internet of things (IoT) [2], people (e.g. experts, skilled workers), and specific domain ontologies. Data and information are maintained and updated on-the-fly by their sources. In order to use data provided by other sources, it seems to be necessary to transform them (at least on format level) to information. For instance, in enterprise engineering domain, as stated in [3], collaborations among various enterprises appear frequently. In some terms, the efficiency of information exchanging (sharing) among heterogeneous partners determines if the collaboration goal could achieve or not. However, this kind of exchanging (sharing) process always involves huge manual work; with the explosion in the amount of data, it is impossible to maintain this process relies mainly on manual work.

Model transformation theories provide a possible solution to share and exchange data/information among

heterogeneous partners. However, there exist several weaknesses in traditional model transformation practices [4]: low reusability, contain repetitive tasks and involve huge manual effort, etc. These weaknesses limit the usage of model transformation theories to serve to cross-domain problems, and also reduce the efficiency of model transformation developing process. For applying model transformation theories to data/information sharing, this paper proposes an automatic model transformation methodology (AMTM) that combines semantic and syntactic checking measurements into model transformation process.

This paper is divided into five sections. The second section presents related work of model transformation domain. Then, a refined model transformation process is stated in the third section. The fourth section presents semantic and syntactic checking measurements in detail. Finally, a conclusion is presented.

2 Related work

This section is divided into two subsections: i) shows the comparisons of several prominent model transformation techniques, and ii) illustrates the category of model transformation practices.

2.1 Model transformation techniques

In this subsection, three popular model transformation techniques are presented and compared briefly.

"ATLAS transformation language (ATL)" [5] is a model transformation language and toolkit. Its architecture composes of three layers: ATLAS Model Weaving (AMW) [6], ATL and ATL Virtual Machine (ATL VM). ATL provides ways to produce a set of target models from a set of source models.

"Query/View/ Transformation (QVT)" [7] is a standard set of languages for model transformation defined by the "Object Management Group"; it covers transformations, views and queries together. The QVT standard defines three model transformation languages. All of them operate on models which conform to Meta-Object Facility (MOF) 2.0 [7] meta-models.

"Visual Automated Model Transformations (VIATRA2)" [8] is a unidirectional transformation language based mainly on graph transformation techniques. The language operates on models expressed following the VPM meta-modeling approach [9].

Table I shows the comparisons on some characteristics among the three techniques.

TABLE I Model Transformation Techniques Comparison

name	hybrid	rule scheduling	M-to-N	note
ATL	yes	implicit internal explicit	yes	self-executed
QVT	no	implicit internal explicit	yes	based on MOF 2.0
VIATRA2	yes	external explicit	yes	graph rewriting

As a short conclusion, many model transformation techniques have been developed. According to the purpose, these techniques could be divided into two groups: serve to "cross-domain" and serve to specific domain.

Normally, domain specific model transformation techniques focus on and provide single solution to particular problematic. The usage of these techniques is limited, and they are not flexible for solving some special cases. On the other hand, cross-domain model transformation techniques provide a wide range of functions, and thus are always complex. So, it needs more time to learn to use this kind of technique properly. The common problem of existing model transformation techniques is: involve huge manual effort and require precondition (e.g. on modeling) to use. To solve this problem, an automatic model transformation methodology that serves to cross-domain, is a possible solution.

2.2 Model transformation category

According to [10], there are two main kinds of model transformation approaches. They are: model-to-code approaches and model-to-model approaches.

As model-to-code approaches, there are two categories: "Visitor-based approaches" and "Template-based approaches". And for model-to-model approaches, there are five categories: Direct-Manipulation Approaches, Relational Approaches, Graph-Transformation-Based Approaches, Structure-Driven Approaches and Hybrid Approaches.

AMTM is a model-to-model model transformation methodology. Based on AMTM, there are two kinds of model transformation situations. Fig. 3 illustrates the two kinds of situations.

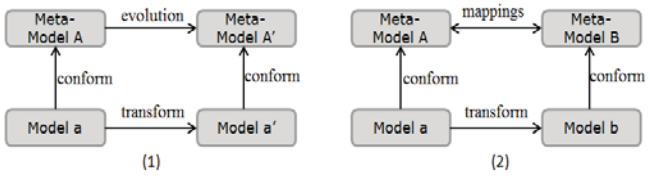


Fig. 3 Model transformation situations

AMTM is created on the basis of meta-model based model transformation methodology.

In situation (a), the target meta-model is the evolutional version of the source meta-model; so, source models that conform to the source meta-model should be transformed to the new version of models that are conformed to the evolutional target meta-model. For this situation, large amount of research work has been done and different theories and practices have been developed. One of the mature theories is "COPE" [11].

In situation (b), source meta-model and target meta-model are created for different purposes. In order to transform source models to target models (conforming to the source and target meta-models, respectively), model transformation mappings should be built on their meta-model level.

AMTM provides a solution to both situation (a) and situation (b); furthermore, there is no precondition of applying AMTM on both of them.

3 Overview of the methodology

This section presents the detail of the automatic model transformation methodology (AMTM). It is divided into two subsections: first subsection illustrates the basic theories of AMTM and second subsection shows the working mechanism of this methodology.

3.1 Basic theories

The basic theories of AMTM are presented in two parts: the theoretical main framework and the meta-meta-model (MMM) involved in it.

3.1.1 Theoretical main framework

AMTM is created on the basis of a theoretical main framework that shows in Fig. 4.

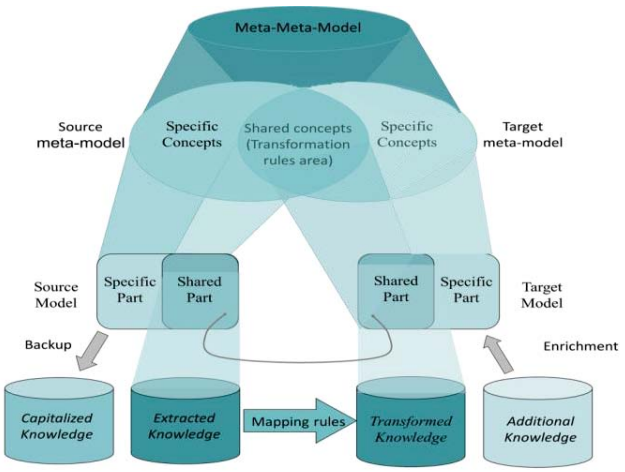


Fig. 4 Theoretical main framework

This theoretical main framework is created based on the work stated in [12]. It illustrates the fundamental theories of AMTM.

For the reason "models are conformed to their meta-models [13]", the potential shared items (between two models) could be traced on meta-model layer. AMTM relies on the meta-model layer (mappings are defined here among shared concepts). The source model embeds a shared part and a specific part. The shared part provides the extracted knowledge, which may be used for the model transformation, while the specific part should be saved as capitalized knowledge in order not to be lost. Then, mapping rules (built based on the overlapping conceptual area between MMs) can be applied onto the extracted knowledge. The transformed knowledge and additional knowledge may be finally used to create the shared part and the specific part of the target model. In order to automatically generate the model transformation mapping rules, semantic and syntactic checking measurements are combined into transforming process (detecting shared concepts on meta-model layer). The principle of applying S&S on model transformation process is stated in [6]. The mechanism of applying S&S in AMTM is defined in the MMM, which shows at the top of Fig. 4.

3.1.2 The meta-meta model

A meta-meta model defines the rules for meta-modeling; there exists several meta-modeling architectures, for example "MOF: Meta-Object Facility" [7]. However, these architectures aim to serve to general purpose; they define their own semantic and syntax. For this project, a specific meta-meta model serves specific to model transformation domain is more preferred. So, within the theoretical main framework, we define this MMM (adapted from MOF) to serve to AMTM. Fig. 5 shows the detail of this MMM.

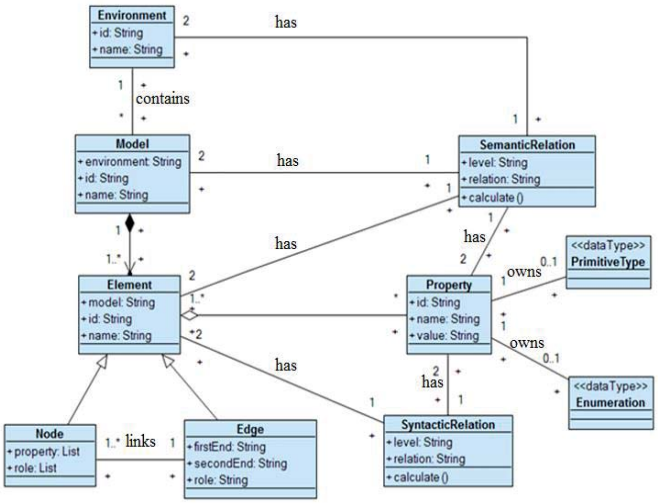


Fig. 5 The meta-meta model

There are ten core elements in this meta-meta-model. As models may come from various domains, a class named "Environment" is defined to stand for these domains. All the model instances are represented by the class "Model", every model belongs to a specific "Environment". "Model" is made of "Element", which has two inheritances: "Node" and "Edge". "Nodes" are linked by "Edge" based on their "roles". "Element" has a group of "Property", the "Property" could identify and explain the "Element". "Property" has a data type: "Primitive Type" or "Enumeration"; to a certain extent, data type could differentiate "Property". Another two key items shown in Fig. 5 are: "Semantic Relation" and "Syntactic Relation". They exist on different kinds of items (e.g. between a pair of elements). Model transformation rules are generated based on these two relations that are existed between different items (i.e. elements and properties).

3.2 AMTM working mechanism

In AMTM, a complete model transformation is regarded as an iterative process: between source model and target model, there could be several intermediate models. An intermediate model could be target model and source model for different transform iterations.

Fig. 6 shows the iterative issue. In each iteration phase, the specific parts from the source meta-model are stored in ontology; the additional knowledge for the specific parts of the target model are enriched by extracting content from the same ontology.

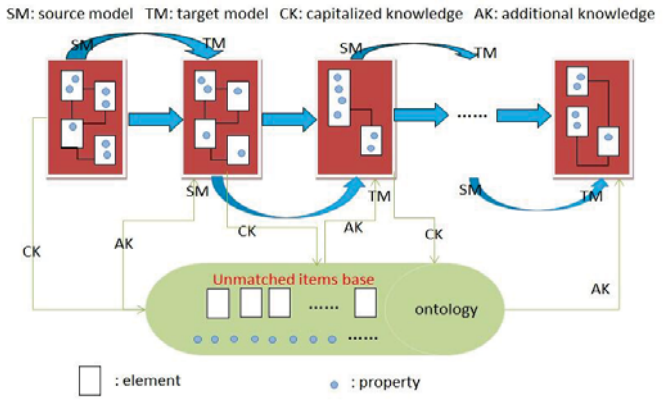


Fig. 6 Iterative model transformation process

To deal with the granularity issue involved, in each iteration, transformation process is divided into three steps: matching on element level, hybrid matching and auxiliary matching.

3.2.1 Matching on element level

According to MMM, meta-models are made of elements. So, model transformation mappings should be defined mainly among elements (nodes and edges); if two elements (come from source model and target model, respectively) stand for the same concept, a mapping should be built between them. The mechanism of defining matching pairs on element level is illustrated by an example shown in Fig. 7.

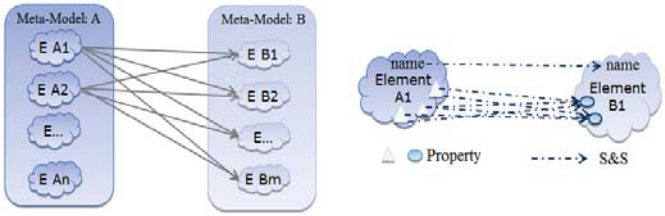


Fig. 7 Matching on element's level

The source meta-model has 'n' elements and 'm' for the target meta-model; the number of comparisons between the two models on element's level is: "m*n".

Within each element's pair, there exists an "Ele_SSV" value. "Ele_SSV" stands for "element's semantic and syntactic value"; it is calculated based on the *elements' names* and their properties. The calculation rule of "Ele_SSV" is shown in (1).

$$\text{Ele\_SSV} = \text{name\_weight}*\text{S\_SSV} + \text{property\_weight} * (\sum_{i=1}^{x} \text{Max}(\text{P\_SSVi}))/x \quad (1)$$

In (1), "name_weight" and "property_weight" are two impact factors for parameters "elements' names" and "elements' properties". Both of the two values are between 0 and 1; the sum of them is 1. "S_SSV" stands for "string semantic and syntactic value; it is calculated based on the words (i.e. element's name). "P_SSV" stands for "semantic and syntactic value between a pair of properties". "x" stands for the number of properties of a specific element from source meta-model (e.g. element E A1).

The example shown below is to calculate the "Ele_SSV" value within the element's pair of "E A1" and "E B1".

"E A1" has number "x" properties and "E B1" has number "y" properties; within each of the "x*y" pairs of properties, there exists a "P_SSV". Equation (2) shows the calculating rule of "P_SSV".

$$\text{P\_SSV} = \text{pn\_weight}*\text{S\_SSV} + \text{pt\_weight}*\text{id\_type} \quad (2)$$

In (2), "pn_weight" and "pt_weight" are two impact factors for the parameters "properties' names" and "properties' types". They play the same role as the impact factors in (1). "S_SSV" is the same as stated in (1); this time, it stands for the semantic and syntactic value between two properties' names. "id_type" stands for "identify properties type". If two properties have the same type, this value is 1; otherwise, this value is 0.

Based on (1) and (2), each element (E A1, E A2…) of the source model gets a maximum "Ele_SSV" with a specific target model element (E B1, E B2…). Moreover, a matching pair of two elements requires building mappings among their properties. The mechanism of choosing matching pairs (on both element and property level) will be illustrated later.

3.2.2 Hybrid matching

After first matching step, some of the elements are still unmatched; even the matched elements, some of their properties are still unmatched. The hybrid matching step focuses on these unmatched items.

This matching step works on property level, all the matching pairs would be built among properties (come from both the unmatched and matched elements). All the unmatched properties from source model will be compared with all the properties from target model. The mechanism of building such matching pairs is also depending on semantic and syntactic checking measurements (based on properties' names and types).

In hybrid matching step, all the matching pairs are built on property's level. *This step breaks the constraint: property matching pairs only exists within matched element's pairs; this constraint is the main granularity issue involved in model transformation process.* However, it is also necessary to consider about the influence from element's level when building mappings in hybrid matching step. The matching mechanism of this step shows in (3).

$$\text{HM\_P\_SSV} = \text{el\_weight}*\text{S\_SSV} + \text{pl\_weight}*\text{P\_SSV} \quad (3)$$

In (3), "HM_P_SSV" stands for "hybrid matching property semantic and syntactic value". "el_weight" and "pl_weight" are two impact factors for the parameters "element level" and "property level". They perform the same functions as the impact factors in former formulas. "S_SSV" is calculated between two elements' names (elements contain the two properties). "P_SSV", as stated in (2), calculates the syntactic and semantic relation between two properties based on their names and types.

3.2.3 Auxiliary matching

After the first and second matching steps, all the shared parts between source model and target model are regarded to be found. However, according to the iterative model

transformation process, there are still some specific parts that should be stored as capitalized knowledge and reused as additional knowledge. This matching step focuses on the mechanism of storing and reusing these specific parts.

All the unmatched items from source model, which regarded as specific parts, are stored in ontology (which is called "AMTM_O" within this project). AMTM_O designed with the same structure as the MMM. For a complete model transformation process, the capitalized knowledge from former iterations could be used as the additional knowledge to enrich the target models that are generated in the latter iterations.

3.2.4 Matching pair choosing mechanism

According to the three former sub-subsections, the relation between two elements is represented by a value between 0 and 1, which calculated by semantic and syntactic comparisons. Based on this value, each element from source model could be matched with "zero to several" elements from target model. The mechanism of selecting elements matching pairs depends on the range of this value. Fig. 8 reveals the basic principle.

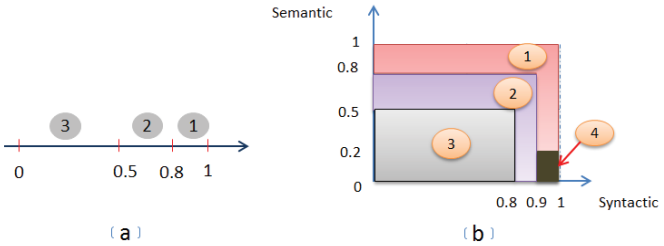


Fig. 8 Matching pair choosing mechanism

For choosing element's matching pairs, two threshold values: 0.5 and 0.8 are assigned. As shown in Fig. 8 situation (a), if two elements have a relation value in region 1 (value between 0.8 and 1), a transformation mapping is built between them; if this value is in region 2, a potential mapping exists between the two elements; else, if this value is in region 3, no mappings will be built between them.

Fig. 8 situation (b) shows the mechanism of choosing matching pairs of two words (i.e. elements' and properties' names). Between two words, strong semantic relation means high potential of making mappings. Region 1 stands for two words that have close relationship: could transform to each other. Region 2 stands for two words have strong relationship: potential transform pair. Region 3 means two words have weak relationship: low possibility to transform to each other. Region 4 is special; it stands for word pairs that have close syntactic relation but very weak semantic relation. For example, word pair: common and uncommon, they could not transform to each other. But in some specific domain (e.g. medicine), syntactic relation may be more important than semantic relation.

In this way, an element (or a property) may have several potential matching items. So, from source model to target model, a "many-to-many" (granularity issue solved in this way) matching relationships are built on both element level and property level.

4 Syntactic & semantic checking

Semantic and syntactic checking measurements play a key role in AMTM. They work together to define a relationship (stands by a value between 0 and 1) between two words. As stated in (1) and (2), the "S_SSV" stands for this value; the calculation method of "S_SSV" is shown in (4).

$$S\_SSV = sem\_weight*S\_SeV + syn\_weight*S\_SyV \quad (4)$$

In (4), "sem_weight" and "syn_weight" are two impact factors for semantic value and syntactic value between two words. The sum of them is 1. "S_SeV" stands for the semantic value between two words, while "S_SyV" stands for the syntactic value.

4.1 Syntactic Checking Measurements

Syntactic checking measurement is used to calculate the syntactic similarity between two words. This kind of checking methodology is based on the alphabets that are contained in the words. Several syntactic checking methodologies have been presented and compared in [14].

The syntactic checking measurements involved in AMTM could be divided into two steps: i) pretreatment: focuses on detecting two words that are in different forms (e.g. tense, morphology) stand for a same word. ii) "Levenshtein Distances" algorithm [15].

For pretreatment, "Porter stemming algorithm" is chosen to be applied in AMTM. "Levenshtein Distances" algorithm is applied between two different words. It calculates the syntactic similarity between two words; the mechanism of using it has been stated in [15].

Equation (5) shows the calculation rule of syntactic relation between two words: word1 and word2, which based on "Levenshtein distances".

$$S\_SyV = 1 - LD / Max(word1.length, word2.length) \quad (5)$$

In (5), "S_SyV" stands for the syntactic similarity value between word1 and word2; "LD" means the "Levenshtein distances" between them. The value of "S_SyV" is between 0 and 1; the higher of this value means the higher syntactic similarity between two words.

4.2 Semantic checking measurements

Different to syntactic checking measurement (relies just on the two comparing words); semantic checking measurement relies upon a huge semantic thesaurus.

A huge semantic thesaurus (AMTM_ST) has been created for serving to AMTM, and AMTM_ST is created on the base of "WordNet" [16]. Fig. 9 shows the structure of AMTM_ST.

As shown in Fig. 9, there are three kinds of items stored in AMTM_ST.

- Word base: contains normal English words (nouns, verbs and adjectives).
- Sense base: contains all the word senses; a word could have "one or several" senses. For example: "star": it has six senses; as noun, it has four senses; as

verb, it has another two senses.

- "Synset" base: a group of word senses that own synonym meanings; semantic relations are built among different synsets.

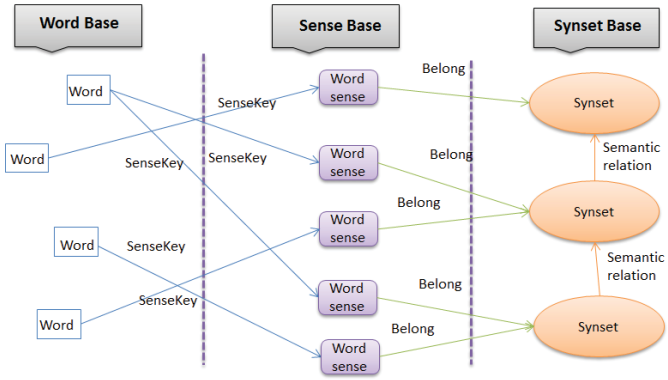


Fig. 9 Structure of AMTM_ST

There are seven kinds of semantic relations that defined among synsets. For each of the semantic relations, a specific value (between 0 and 1) is assigned to it. Table II shows these semantic relations and their corresponding values.

TABLE II SEMANTIC RELATIONS AND VALUES IN AMTM_ST

Semantic relation	S_SeV	Remark	Example
synonym	0.9	words from the same synset	shut & close
hypernym	0.6	two synsets have this relation	person-creator
hyponym	0.8	two synsets have this relation	creator-person
similar-to	0.85	only between two adjectives	perfect & ideal
antonym	0.2	words have opposite meanings	good & bad
iterative hypernym	0.6^n	iterative hypernym relation	person-creator-maker-author
iterative hyponym	0.8^n	iterative hyponym relation	author-maker-creator-person

In table II, all the "S_SeV" values are assigned directly (based on experience); these values could be assigned with different values by different application domains.

As a word may have different word senses (furthermore, may belong to different synsets), there might be several semantic relations that exist between two words. So, the number of "S_SeV" values between two particular words is not limited to one. In this project, we focus on finding the maximum "S_SeV" value between two words. Based on this cognition, the process of detecting semantic relations between two words should be serialized.

In order to define the semantic relation between two words, there are several steps to follow:

- First, locating two words (element's or property's names) in AMTM_ST.
- Second, finding all the word senses of the two words and grouping these word senses into two sets.
- Third, tracing all the synsets, which the two sets of word senses belong to, and grouping these synsets into two groups.

After getting two synsets groups, the final step is to detect the semantic relations that exist among all the possible synset pairs (one from word1 side, the other from word2 side). For detecting five kinds of semantic relations: synonym, similar-to, hypernym, hyponym and antonym, the basic principle is: search all the synsets that have these five kinds of semantic relations with the synsets in "synset group of word1", then comparing if there exist one synset in "synset group of word 2", which is the same as one of the located synsets.

The detecting process of "iterative hypernym" and "iterative hyponym" semantic relations is same. The basic idea is: locating the synsets that have hypernym relation with word1's synsets iteratively and comparing with word2's synsets, in order to find two same synsets.

The basic information of doing all these semantic checking measurements is provided by AMTM_ST. So, the content of AMTM_ST should be really huge. For serving AMTM, AMTM_ST contains 147306 words, 206941 word-senses and 114038 synsets.

4.3 Short conclusion

By using syntactic and semantic checking measurements, a "S_SSV" value could be generated between two words. When the words stand for properties' names, an approximate value between two properties is generated (properties' types are also considered). When the words stand for elements' names, an approximate value between two elements is generated (the summary of approximate values on their properties' level is also considered). Based on all these approximate values, model transformation mappings could be built automatically between source and target models.

5 Conclusions

In this paper, an automatic model transformation methodology (AMTM) is presented. According to the real requirement "exchanging information effectively and efficiently", model transformation should be done automatically. So, semantic and syntactic checking measurements are combined into model transformation process to replace manual effort.

As theoretical foundation, a main framework is created; within this framework, a meta-meta-model is defined to present the mechanism of combining semantic and syntactic checking measurements into the process of defining model transformation mappings. For syntactic checking, "Porter stemming algorithm" and "Levenshtein distance algorithm" are used. For the semantic checking measurement, a specific semantic thesaurus (AMTM_ST) is built. To deal with the granularity issue, model transformation is regarded as an iterative process and within each iteration phase, the transformation process is divided into three steps. Furthermore, a specific ontology (AMTM_O) is created to support the third matching step: auxiliary matching. This AMTM_O helps to store specific parts from source models and enrich specific parts for the target models.

However, there are some points in this AMTM that needed to be improved in the future.

- The impact factors such as "sem_weight", "pn_weight" and threshold values for choosing matching pairs: the better way to assign them is "using some mathematic strategy" (e.g. "choquet" integral; one of the usages of "choquet" integral is stated in [17]).
- Semantic checking measurement: only formal English words are stored in the semantic thesaurus; for words that in specific cases or phrases, they cannot be located in AMTM_ST.
- The S_SeV values that defined in table II: they should be assigned differently based on the specific application domains.
- Matching pair choosing mechanism: we aim at finding the strongest semantic relation between two words, but the chosen semantic meaning may not be the exact one that the words conveyed within a specific context.

Furthermore, the usage of AMTM is cross-domain; AMTM aims at transforming and combining rough data to information (with specific structure and format), and then exchanging information among different domains.

Fig. 10 shows the scientific contribution of AMTM: converting rough data to information and exchanging and merging information on the information platform.

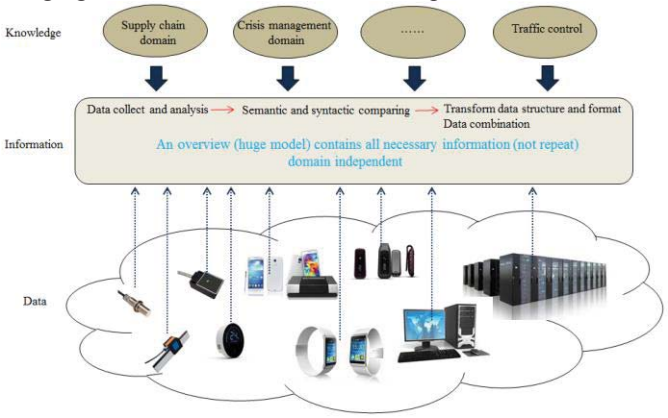


Fig. 10 AMTM scientific contribution

Many data collectors (IoT) such as: sensors, smart equipment, computers, could gather rough data from a particular region or domain. Generally, this kind of data focuses on different purposes and reflects different views of a system. Moreover, different collectors store data in heterogeneous structures. AMTM regards these collected data as many single models, and uses semantic and syntactic checking measurements to detect the intrinsic links among them. Finally, after transforming and combining these data, a huge model (overview of a specific system) is generated. This huge model contains all the useful (not overlap) information. With rules that defined in specific domains, such information could be transformed (exchanging & sharing) to knowledge which serves to domain specific problems.

By combining semantic and syntactic checking measurements into model transformation process, an efficient model transformation methodology "AMTM" is created. With the improvement on some detail aspects, this methodology could serve to information sharing issue for a large number of domains in practice.

6 References

[1] R.L. Ackoff, "From Data to Wisdom", journal of applied system analysis, volume 16, 1989: 3-9.

[2] L. Atzori, A. Iera, G. Morabito, The Internet of Things: a survey Computer Networks, 54 (2010), pp. 2787–2805.

[3] LM. Camarinha-Matos and H. Afsarmanesh, 'Classes of Collaborative Networks', in Encyclopedia of Networked and Virtual Oragnization, ed. by Goran D. Putnik and Maria Manuela Cunha, Information Science Reference (Hershey, New York, 2008), I, 193–98.

[4] M.D. Del Fabro, P. Valduriez, Towards the efficient development of model transformations using model weaving and matching transformations. Software & System Modeling, July 2009, Volume 8, Issue 3, pp 305-324.

[5] F. Jouault, F. Allilaire, J. Bézivin, I. Kurtev, ATL: A model transformation tool. Science of Computer Programming. 2007, Volume 72, Issues 1–2.

[6] M.D. Del Fabro, J. Bézivin, F. Jouault, E. Breton, AMW: A Generic Model Weaver. 2005, 1ère Journées sur l'Ingénierie Dirigée par les Modèles: Paris.

[7] OMG: QVT. Meta Object Facility (MOF) 2.0 Query/View/ Transformation Specification OMG (2008) http://www.omg.org/spec/QVT/1.0/PDF

[8] D. Varr´o, A. Balogh, The model transformation language of the viatra2 framework. Sci. Comput. Program 68(3), 214–234 (2007).

[9] D. Varr´o, A. Pataricza, VPM: A visual, precise and multilevel meta-modeling framework for describing mathematical domains and UML (the mathematics of metamodeling is metamodeling mathematics), Software. Syst. Model. 2 (3) (2003) 187–210.

[10] K. Czarnecki, S. Helsen, Classification of Model Transformation Approaches. OOPSLA'03, 2003, Workshop on Generative Techniques in the Context of Model-Driven Architecture.

[11] M. Herrmannsdoerfer, S. Benz, E. Juergens, COPE - automating coupled evolution of metamodels and models. In: Drossopoulou, S. (ed.) ECOOP 2009 – Object-Oriented Programming. LNCS, vol. 5653, pp. 52–76. Springer, Heidelberg.

[12] F. Bénaben, W. Mu, S. Truptil, H. Pingaud, "Information Systems design for emerging ecosystems." 2010, 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST).

[13] J. Bezivin, "Model driven engineering: An emerging technical space," in Generative and Transformational Techniques in Software Engineering, International Summer School -GTTSE, 2006, pp. 36–64.

[14] W. C. William, R. Pradeep, E. F. Stephen, "A Comparison of String Metrics for Matching Names and Records." KDD Workshop on Data Cleaning and Object Consolidation, 2003, Vol. 3.

[15] H. Wilbert, "Measuring Dialect Pronunciation Differences using Levenshtein Distance." Ph.D. thesis, 2004, Rijksuniversiteit Groningen.

[16] X. Huang, C. Zhou, "An OWL-based WordNet lexical ontology." Journal of Zhejiang University, 2007, pp. 864-870.

[17] D. Abril, G. Navarro-Arribas, V. Torra, Choquet integral for record linkage Ann. Oper. Res., 195 (1) (2012), pp. 97–110

Knowledge Representation: Conceptual Graphs vs. Flow-Based Modeling

Sabah Al-Fedaghi
Computer Engineering Department, Kuwait University, Kuwait

Abstract - *One of the basic principles of knowledge representation is that it is a language by which people say things about the world. Visual depictions appear particularly useful for representation of knowledge, e.g., Peirce's Existential Graphs, and Sowa's Conceptual Graphs (CGs). Recently, a new flow-based model for representing knowledge, called the Flowthing Model (FM), has been proposed and used in several applications. This paper is an exploratory assessment of the capability of FM to express knowledge, in contrast to CGs. Initial examination suggests that FM contributes to* expressing *knowledge in a way not provided by CGs. In addition, FM seems to produce a new aspect that may complement the CG formalism. Such exploration can promote progress in knowledge representation and modeling paradigms and their utilization in various applications.*

Keywords: Knowledge representation; conceptual graphs; diagrammatic representation; flow-based knowledge representation

1 Introduction

During the past 40 years, visual depictions have been used in the area of knowledge representation, specifically using a semantic network, e.g., [1-2]. Many of these representations concentrate on the fields of linguistic knowledge (e.g., [3]), knowledge in large-scale development of applications, or logical aspects of semantic networks. A basic principle of knowledge representation, as a medium of human expression, is that it is "a language in which we say things about the world" [4]. This paper focuses on this aspect of knowledge representation, in contrast to such features as logical reasoning or computational efficiency.

To limit the scope of this paper, it examines Sowa's Conceptual Graphs (CGs) and contrasts them with a newly developed conceptual representation based on the notion of *flow*. The CG was developed as "a graph representation for *logic* based on the semantic networks of artificial intelligence and Peirce existential graphs" [5; italics added]. Initially, Sowa developed CGs as an intermediate language for mapping natural language assertions to a relational database. They have also been viewed as a diagrammatic system of logic to express meaning in a precise form, humanly readable, and computationally tractable [5].

CGs have been applied in a wide range of fields [5]. In artificial intelligence, CG formalism offers many benefits, including graph-based reasoning mechanisms, plug-in capabilities over data structures, and good visualization capabilities [6]. In addition, a conceptual graph can serve as an intermediate language for translating computer-oriented formalisms to and from natural languages [7]. They provide a readable but formal design and specification language.

> The research CGs have explored novel techniques for reasoning, knowledge representation, and natural language semantics. The semantics of the core and extended CGs is defined by a formal mapping to and from ISO standard 24707 for Common Logic, but the research CGs are defined by a variety of formal and informal extensions. [5]

CGs are an intuitive, visual way of creating a semantically sound representation of knowledge [8], and many extensions have been proposed (e.g., [9-10])

Recently, a new flow-based model for representing knowledge, called the Flowthing Model (FM), has been proposed and used in several applications, including communication and engineering requirement analysis [11-15]. This paper is an exploratory assessment of the capability of FM to express knowledge in the domain of CGs. Initial examination has suggested interesting results when CGs and FM-based diagrams are drawn to depict the same representation. FM seems to produce a new aspect that may complement the CG formalism.

CGs have a solid foundation, not only in the area of knowledge representation but also in reasoning. It is a well-known formalism that hardly needs describing. FM is still an informal description, but it introduces an interesting high-level schematization of knowledge-related problems. It is based on the notion of flow and includes exactly six stages (states) where "things" can transform from one stage to another in their life cycles. The paper examines CGs and FM to contrast common concepts and differences between the two methodologies. Several advantages can be achieved from such a study:

- Enhancing conceptualization of the common concepts by modeling them from completely different perspectives
- Benefiting the foundation of knowledge representation by subjecting the same piece of knowledge to two dissimilar views of modeling

• Enhancing and complementing of each approach by the other, promoting progress in knowledge representation and modeling paradigms and their utilization in various applications.

For the sake of completeness, and because FM is not yet a well-known methodology, the model is briefly described in the next section. The example presented in the section is a new contribution.

2 Flowthing Model

A flow model is a uniform method for representing things that "flow," i.e., things that are created, processed, released, transferred, and received. "Things that flow" include information, materials (e.g., goods), and money. They flow in *spheres*, i.e., their environments. A sphere is different from a set in the sense that a set is a static structure, whereas a sphere includes flowthings (current members) at different stages in a progression and possible directions (lines) of movement from one stage to another, or movement from/to the spheres of the flowthings. A sphere may have subspheres.

An FM representation is a depiction of the structure of a scheme resembling a road map of components and conceptual flow. A *component* comprises *spheres* (e.g., those of a company, a robot, a human, an assembly line, a station) that enclose or intersect with other spheres (e.g., the sphere of a house contains rooms which in turn include walls, ceilings). Or, a sphere embeds flows (called *flowsystems*; e.g., walls encompass pipes of water flow and wires of electrical flow).

Things that flow in a flowsystem are referred to as *flowthings*. The life cycle of a flowthing is defined in terms of six mutually exclusive *stages*: creation, process, arrival, acceptance, release, and transfer.

Fig. 1 shows a flowsystem with its stages, where it is assumed that no released flowthing flows back to previous stages. The reflexive arrow in the figure indicates flow to the Transfer stage of another flowsystem. For simplicity's sake, the stages Arrive and Accept can be combined and termed *Receive*.

The *stages* of the life cycle of a flowthing are mutually exclusive (i.e., a flowthing can be in one and only one stage at a time). All other states or conditions of flowthings are not exclusive stages. For example, we can have *stored* created flowthings, *stored* processed flowthings, *stored* received flowthings, etc.; thus *stored* is not a specific stage. In contrast, there are no such stages as, e.g., *created-processed*, *received-transferred*, or *processed-received* stages. Flowthings can be released but not transferred (e.g., the channel is down), or arrived but not accepted (wrong destination), …

In addition to flows, *triggering* is a transformation (denoted by a dashed arrow) from one flow (or stage) to another, e.g., a flow of electricity triggers a flow of air.

Example: In 1883, Peirce [16] developed a graphical notation for logical expressions called an existential graph (EG). Fig. 2 shows this graph for the statement *You can lead a horse to water, but you can't make him drink* [17]. Fig. 3 shows its corresponding FM representation.

In Fig. 3, there are three spheres: Horse, Water place, and Person. A person *creates* an action that triggers the *transfer* of the horse to the water place. In the water place, the horse is *received* and *processed*. Process here refers to doing something to the horse, in this case, trying to make the horse drink, but the end result is that the horse does not drink.

It is possible to model that the person actually tries to *force* the horse to drink by *triggering* another action (e.g., actually forcing the horse's mouth into the water; see Fig. 4). *Not drinking* indicates that whatever "processing" was used on the horse (e.g., getting it wet), the horse did not drink.

As can be seen, it is difficult to present a *technical* comparison of two diagramming methodologies; however, putting them side by side allows for a visually mediated awareness of the features and expressive nature of each methodology. This general approach is adapted in this paper regarding CGs and FM. Of course, this analysis does not extend to the reasoning capability embedded in CGs.

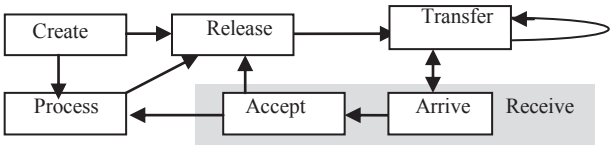


Fig. 1. Flowsystem

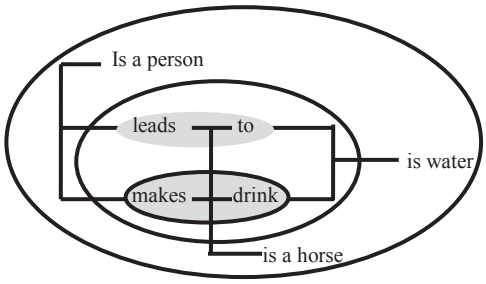


Fig. 2. EG representation of *You can lead a horse to water, but you can't make him drink.*

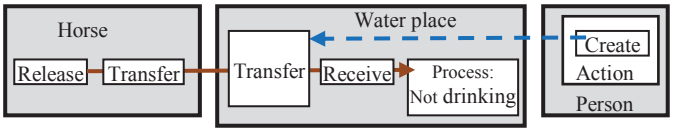


Fig. 3. FM representation of *You can lead a horse to water, but you can't make him drink.*

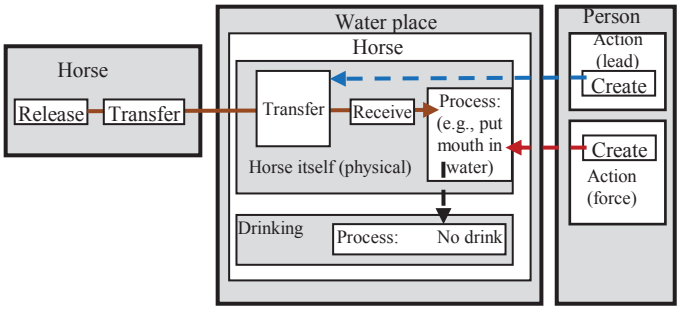


Fig. 4. Another interpretation of *You can lead a horse to water, but you can't make him drink.*

3 FM and CGs

This section contrasts conceptual graphs and FM by re-casting several CG diagrams in FM representations.

A. Example 1

In CGs, *functions* are represented by conceptual relations called actors. Fig. 5 shows the CG for the equation

$$y = (x + 7) / sqrt(7)$$

According to Sowa [5], the equation would be represented by three actors (diamond-shaped nodes). "The concept nodes contain the input and output values of the actors. The two empty concept nodes contain the output values of Add and Sqrt" [5].

Fig. 6 shows the corresponding FM representation. We assume that *x* is input and 7 is a stored (available) constant in the manner of programming languages. Accordingly, *x* and 7 are received, each in its own flowsystem, and both flow (3) to *Add* (4). In *Add*, they are processed (added) to trigger the creation of a sum that flows to *Divide*. On the other hand, 7 flows to *Sqrt* to be processed to create a square root that also flows to *Divide*. In *Divide*, the division operation is performed (which term is divided by which can be specified beforehand in the process stage), to *Create y* as output.

Contrasting the two representations from a style point of view, it appears that CG uses extra constructs (represented as shapes in the graph) accompanied by a potentially infinite number of verbs, e.g., add, divide, subtract, … while FM utilizes the notion of flowsystem, repeatedly using the generic six stages.

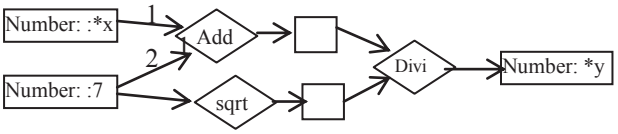


Fig. 5. Functions represented by actor nodes (from [5])

B. Example 2

Mineau et al. [8] present an example of the text of instructions for decalcifying a coffee machine that can be represented by CG (Fig. 7; see their source of instructions):

> In order to decalcify a coffee machine in an environment friendly way, one must fill it up with water and put in two teaspoons of citric acid (from the drugstore). Then, one must fill it up with clear water and let it go through the machine twice. [8]

Fig. 8 shows the corresponding FM representation, according to what is understood from the original description. First, two teaspoons of citric acid (circle 1) are added to water (2) to trigger (3) the creation of a mixture (4). The mixture is poured into the machine (5), followed by triggering (6) the state of the machine to be ON (7). Note here that triggering indicates a control flow. After that the mixture is processed (7) and then released to the outside (8), followed by triggering (9) of pouring water into the machine (10) - twice.

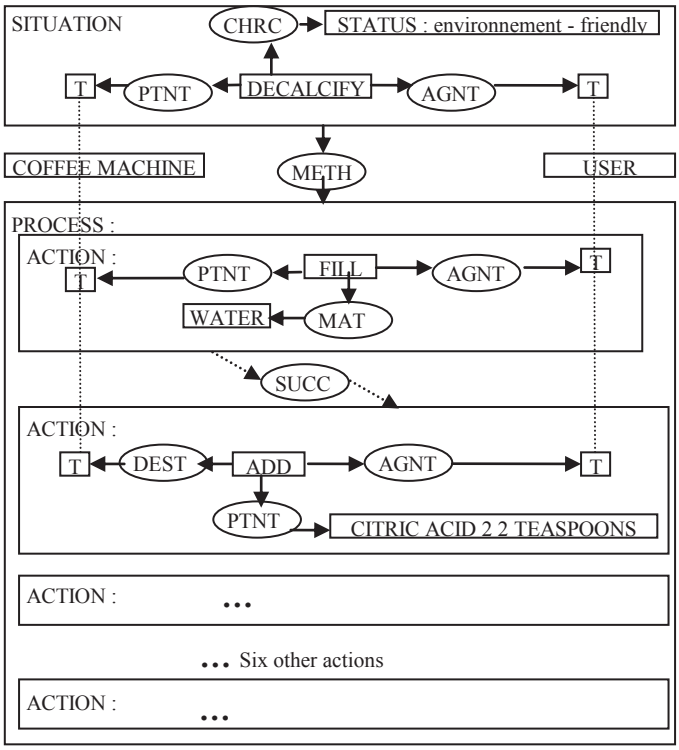


Fig. 7. CG representation of decalcifying (partial, from [8])

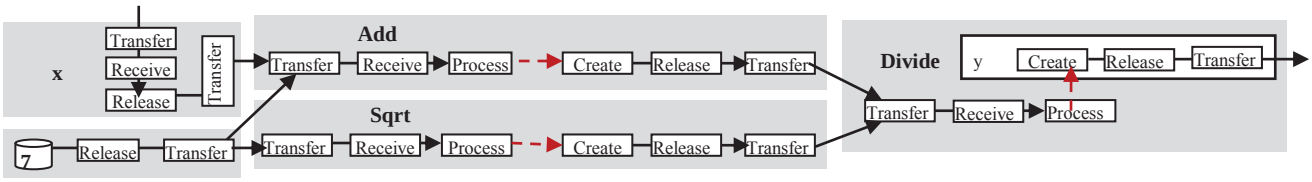


Fig. 6. FM representation of the function

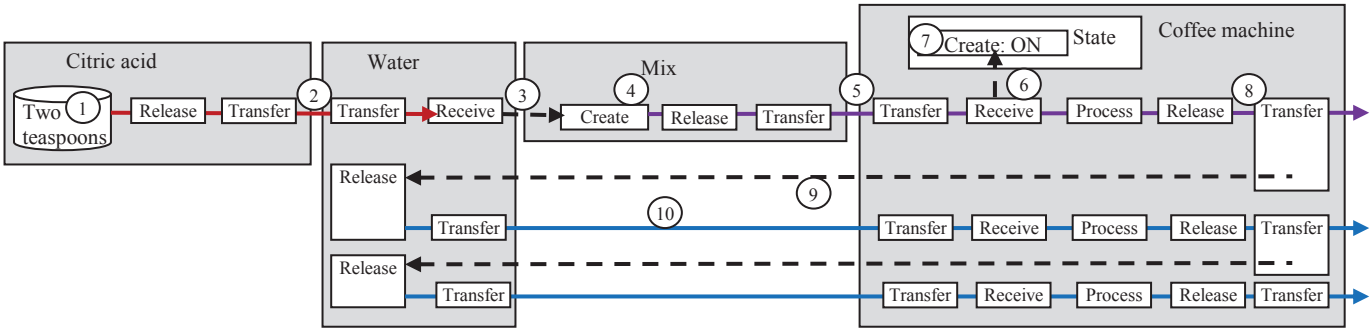


Fig. 8. FM representation of decalcifying

Again, it is difficult to present a *technical* comparison between the two diagramming methodologies (e.g., showing number of nodes, their shapes, edges, …). However, putting the diagrams side by side provides the reader a sense of the expressiveness or understandability of each method, since, as stated by [4], knowledge representation is "a language in which we say things about the world."

C. Example 3

According to [5], the most common use of language is to talk about beliefs, desires, and intentions. As an example, the sentence *Tom believes that Mary wants to marry a sailor*, contains three clauses, whose nesting can be indicated by brackets:

Tom believes that [Mary wants [to marry a sailor]]

The outer clause asserts that *Tom has a belief*. Tom's belief is that *Mary wants* a situation described by the nested infinitive. Each clause makes a comment about the clause or clauses nested in it. The original sentence can be interpreted as,

Tom believes that [there is a sailor whom Mary wants [to marry]]. That is, there is a sailor whom Tom believes that [Mary wants [to marry]]. Fig. 9 shows these interpretations using CGs with case relations.

Fig. 10 shows a corresponding FM representation of the second interpretation. *Tom* (who is an existing person, as indicated by the ability to *Create*) has an existing (Create) belief (box inside *Tom*) in a *Proposition* (idea existing in his mind – *Create*). The inner part of the figure (box drawn with thick lines, blue in the online version) is a description of the proposition. It does not have *Create* because it is the schematic portrait of the proposition, the same way a blueprint of the house is not the house itself. The proposition is about *Mary*, who has a desire: *being married to a sailor*.

Now, suppose that Mary and the sailor are existing persons. Then the mapping between "reality" and what's in Tom's mind is shown in Fig. 11. In the world of Fig. 11, there are three persons (spheres): *Tom* (1), *Mary* (2), and the *Sailor* (3). The "real" *Mary* (2) and *Sailor* (3) trigger images (concepts) of *Mary* (4) and *Sailor* (5) in Tom's mind (we could have a box for *Mind* in Tom's sphere, but this is implicitly understood). Thus, *Mary* (7) and *Sailor* (8) in the proposition "refer to" (trigger) these images of Mary and Sailor (5 and 6). Apparently, the "real" *Mary* (2) did something (9) that triggered the creation of a belief in *Tom's* mind (10).

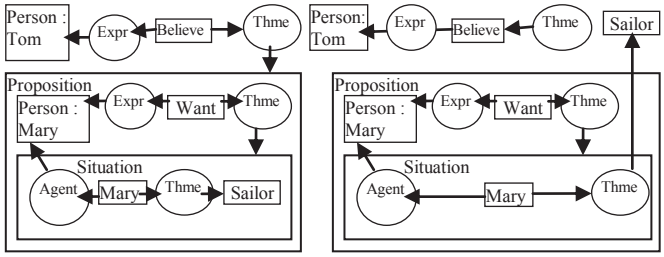


Fig. 9. Two CG interpretations of *Tom believes that Mary wants to marry a sailor*

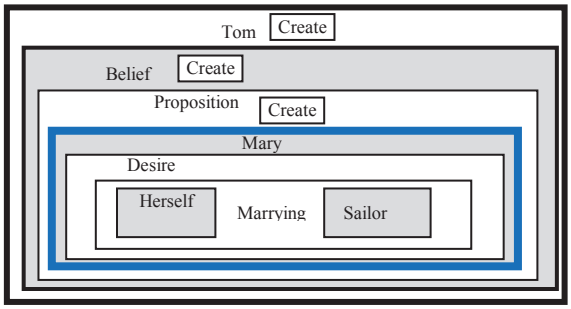


Fig. 10. FM representation of *Tom believes that Mary wants to marry a sailor*

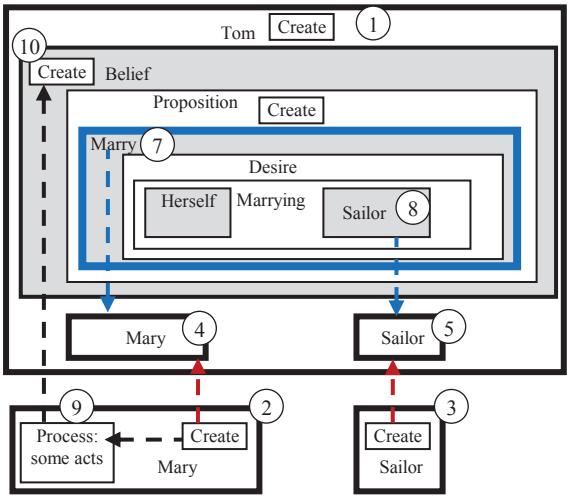


Fig. 11. FM representation of *Tom believes that Mary wants to marry a sailor*, interpreting Mary and Sailor as real persons

Now, suppose that there is a sentence in "reality" (e.g., published in a magazine) that *Tom believes that Mary wants to marry a sailor*; accordingly, this sentence, as shown in Fig. 12, exists (is created, circle 11). Furthermore, suppose that Tom himself first expressed that sentence. Accordingly, Tom's belief is converted into a mental linguistic "sentence" (Fig. 12, circle 12), *I believe that Mary wants to marry a sailor*, that is triggered (13) by his belief.

The purpose of these assumptions is to demonstrate certain aspects of FM. The resultant FM "knowledge" representation seems to be a clear map of the various "items" involved in the situation: *Tom*, *Sentence*, *Mary*, and the *Sailor* have different "realizations" according to the sphere (e.g., Mary in reality, in Tom's mind, and in a linguistic expression). Inside Tom, we find the "concepts" of Mary, Sailor, his Belief, and his Sentence. Within that belief, we find the proposition and its "meaning." Comparing the FM (Fig. 12) with the CG (Fig. 9) with its two interpretations, it seems that the FM description provides "something" that has not been captured by the CG.

D. Example 3

Sowa [5] gives a CG for the sentence *John is going to Boston by bus* (see Fig. 13(a)). The rectangles represent concepts, and the circles represent conceptual relations. "An arc pointing toward a circle marks the first argument of the relation, and an arc pointing away from a circle marks the last argument" [5]. There are three conceptual relations: agent (Agnt), destination (Dest), and instrument (Inst). Fig. 14 shows the corresponding FM representation. In Fig. 13(b), John *flows* to the bus, and after he gets on the bus, the bus flows to Boston.

According to Sowa [5], The CG can be translated to the following formula:

(∃x)(∃y)(Go(x) ∧ Person(John) ∧ City(Boston) ∧ Bus(y) ∧ Agnt(x,John) ∧ Dest(x,Boston) ∧ Inst(x,y))

But such a formula introduces new information, that John and the bus *exist* (e.g., the sentence is not song lyrics). Boston's existence is implied implicitly (no (∃z) and Location (z, Boston)). Accordingly, we modify the FM representation by making *John* and *Bus* exist (Create), as shown in Fig. 15.

Comparing the two representations, it seems that CG needs additional "semantics" (case relations) by trying to apply roles to concepts (e.g., agent, destination, and instrument), however, this is not used uniformly, e.g., Boston is a location. FM looks more uniform and "simple."

Additionally, the FM representation of Fig. 13(b) may be susceptible to logic formulas, e.g.,

(∃x)(∃y)(flow(x, y) ∧ flow(y, Boston) ∧ Is(x, John) ∧ Is(y, Bus)

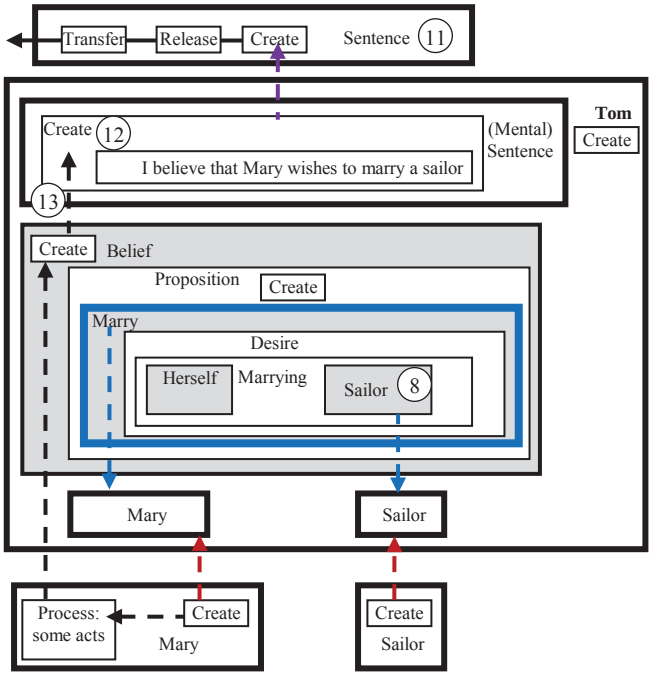


Fig. 12. FM representation that includes the sentence *Tom believes that Mary wants to marry*

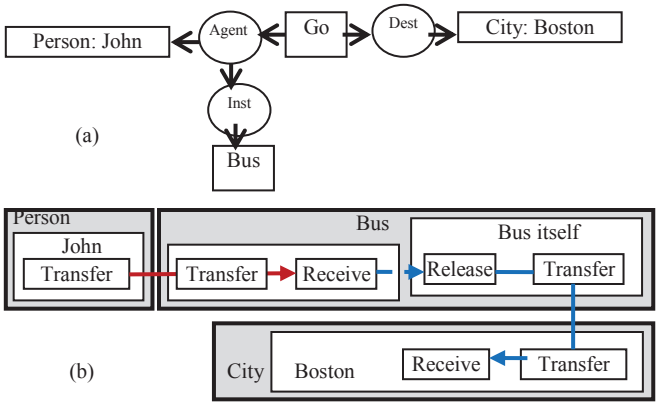


Fig. 13. *John is going to Boston by bus* in (a) CG and (b) FM

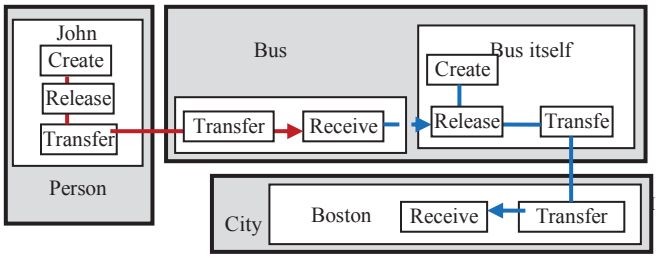


Fig. 14. *John is going to Boston by bus* in FM as supplemented by the logic formula

4 Expressing Constraints

This section applies the FM model to specify constraints in a known CG-based case study. In FM, the constraint is an integral part of the flow-based diagrammatic model.

The Sisyphus-I case study [18] is a well-known resource allocation problem where it is required to allocate the members of a research group to different offices, given certain constraints. "Constraints are used to verify the validity of worlds (the world description, enriched by implicit knowledge rules, must satisfy every constraint)" [19]. For constraints, an intuitive semantic could be *if information A is present, then we must also find information B* [19].

Fig. 15 (from [19]) shows a sample constraint representation in pure CG (using graphs with colored nodes). Fig. 16 shows the corresponding FM representation. The FM representation seems to be simpler since "in" is implicit in the diagram and there is no need to have colored nodes. Note that "in" in Fig. 15 can be interpreted in different ways. It may mean the assignment of the office as a "place of work." In this case, Fig. 17 expresses that the office is where the boss or the person does his or her work. Or, we can merge the two interpretations of Figs. 16 and 17, and consider "in" to mean the office that takes (receives) the boss when he/she comes to the company and where the boss works.

Consider another constraint described in the literature of the Sisyphus-I case study: *If offices are shared, smoking preference should be the same. In other words, smokers and non-smokers should not be allocated to the same room* [20]. Fig. 18 shows its CG representation. The figure is not complete because the purpose here is not to present a complete and fair description of the constraint and its representation; rather, the aim is to show the type of diagramming method used and to contrast it with the FM description.

Fig. 19 shows the corresponding FM diagram, drawn according to our understanding of the involved constraint. In Fig. 19, the sharing sphere (circle 1) has two types of offices: smoking and nonsmoking (2 and 3). If there are two persons in any of these offices, say, x and y, then their preferences are the same. Note that x and y are spheres that have two sub-spheres, *Work* and *Preference*. *Process* in the subsphere (e.g., 4) indicates that the person is performing work (e.g., not a visitor). *Create* (e.g., 5) in the office sphere indicates that the preference is mandatory.

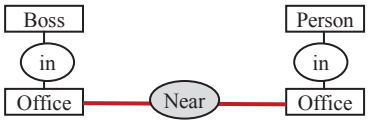


Fig. 15. Pure CG for the constraint *if a boss is in an office and a person is in an office, then this latter office must be near the boss's office*

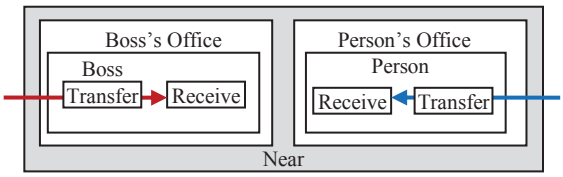


Fig. 16. FM representation of the constraint

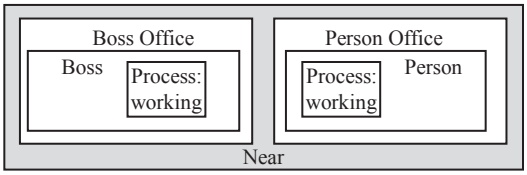


Fig. 17. FM representation of the constraint

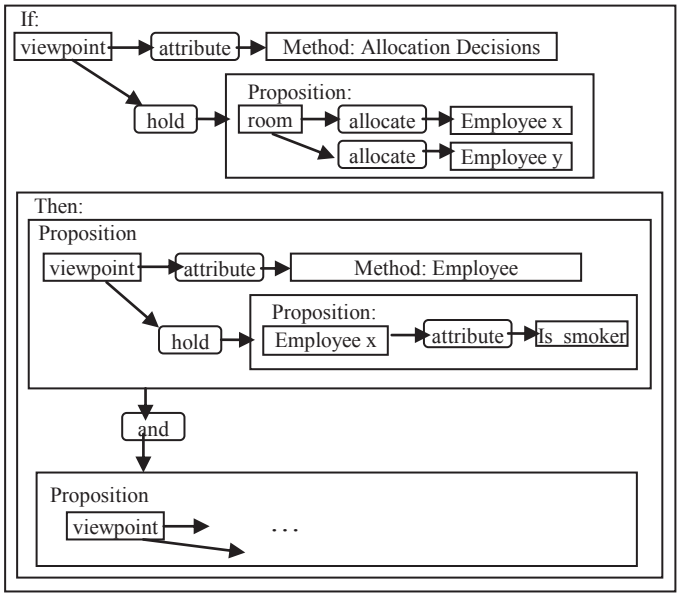


Fig. 18. CG (with case relations) representation of *If offices are shared, smoking preference should be the same. In other words, smokers and non-smokers should not be allocated to the same room* [partial, redrawn from [20]).

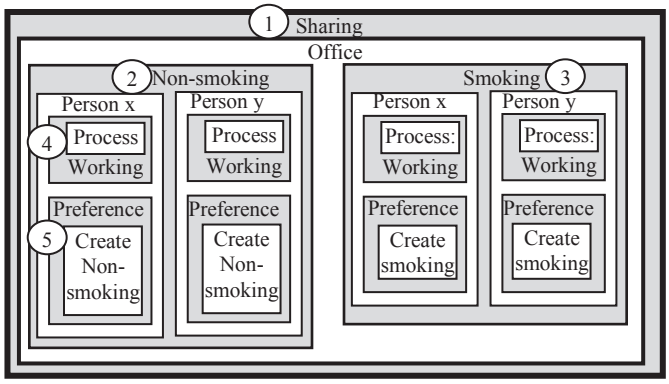


Fig. 19. FM representation of sharing an office

5 Conclusions

This paper focuses on two approaches to diagramming knowledge representation: Conceptual Graphs (CG), and a newly proposed diagramming methodology (FM) based on the notion of flow. The aim is to contrast common concepts and differences between the two methodologies and implicitly to raise the issue that FM may contribute to *expressing* knowledge in a way that is not provided by CGs. There is also the possibility that FM can be applied in reasoning, but this particular issue is not considered in the paper.

This paper is an exploratory assessment of the capability of FM to express knowledge in the domain of CG. By comparing examples, this initial examination seems to suggest the viability of FM-based diagrams to depict at least a high-level representation that is different in role and application in comparison with CGs. There is also the possibility of a new aspect that may complement the CG formalism. Further research will further explore such issues.

6 References

[1] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and Pe.F. Patel-Schneider, The Description Logic Handbook: Theory, Implementation, and Applications. New York: Cambridge University Press, 2003.

[2] R.J. Brachman and J.G. Schmolze, "An overview of the KL-ONE representation system," Cognitive Sci, vol. 9(2), pp. 171-216, 1985.

[3] R.C. Schank and C.J. Rieger, "Inference and computer understanding of natural language," Artif. Intell., vol. 5, pp. 373-412, 1974.

[4] R. Davis, H. Shrobe, and P. Szolovits, "What is a knowledge representation?" AI Mag., vol. 14(1), pp. 17-33, 1993.

[5] J.F. Sowa, "Chapter 5: Conceptual graphs," in Handbook of Knowledge Representation, F. van Harmelen, V. Lifschitz, and B. Porter, Eds. Elsevier, 2008, pp. 213-237. http://www.jfsowa.com/cg/cg_hbook.pdf

[6] M. Croitoru, Conceptual graphs at work: efficient reasoning and applications. Ph.D. diss., University of Aberdeen, 2006.

[7] M. Obitko, Introduction to Ontologies and Semantic Web: Translations between Ontologies in Multi-Agent Systems. Ph.D. diss., Faculty of Electrical Engineering, Czech Technical University, Prague, 2007. http://www.obitko.com/tutorials/ontologies-semantic-web/conceptual-graphs.html

[8] G.W. Mineau, "A first step toward the knowledge web: Interoperability issues among conceptual graph based software agents, Part I." LNCS 2393, pp. 250-260. Springer, 2002.

[9] M. Willems, "A conceptual semantics ontology for conceptual graphs," in Proceedings of the 1st International Conference on Conceptual Structures (ICCS'93), pp. 312-327, 1993.

[10] M. Croitoru, E. Compatangelo, and C. Mellish, "Hierarchical knowledge integration using layered conceptual graphs," in Proceedings of the 13th Internatonal Conference on Conceptual Structures (ICCS'2005), no. 3596 in LNAI, pp. 267-280. Springer, 2005.

[11] S. Al-Fedaghi, "Schematizing proofs based on flow of truth values in logic," IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), Manchester, UK, 2013.

[12] S. Al-Fedaghi, "Visualizing logical representation and reasoning." The 15th International Conference on Artificial Intelligence (ICAI'13), July 22-25, 2013, Las Vegas, USA.

[13] S. Al-Fedaghi, "Enriched diagrammatic representation for resolution," Asia Pacific Symposium on Intelligent and Evolutionary Systems, November 7-9, 2013, Seoul, Korea. Also appears in Procedia Computer Science, Elsevier.

[14] S. Al-Fedaghi, "How to diagram your drama story." HCI International 2013, July 21-26, 2013, Las Vegas, Nevada, USA. In Lecture Notes in Computer Science/Artificial Intelligence, Springer. Extended version: Communications in Computer and Information Science (CCIS) series, vol. 374, 2013, pp 531-535. Springer.

[15] S. Al-Fedaghi, "High-level representation of time in diagrammatic specification," Procedia Comp. Sci. J. Paper presented at 2015 International Conference on Soft Computing and Software Engineering (SCSE'15), University of California at Berkeley (UC Berkeley), USA, March 5-6, 2015.

[16] C.S. Peirce, Manuscript 514, 1909. Available at http://www.jfsowa.com/peirce/ms514.htm.

[17] J.F. Sowa, Semantic Foundations of Contexts. http://users.bestweb.net/~sowa/ontology/contexts.htm

[18] Sisyphus'94. Int. J. Human-Comp. Studies, Special Issue: Sisyphus: Models of Problem Solving, vol. 40(2), 1994.

[19] J.-F. Baget, D. Genest, and M.-L. Mugnier, "Knowledge acquisition with a pure graph-based knowledge representation model: Application to the Sisyphus-I case study," in Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99).

[20] T. Thanitsukkarn and A. Finkelstein, "Multiperspective analysis of the Sisyphus-I room allocation task modelled in a CG meta-representation language," in Conceptual Structures: Standards and Practices, W. Tepfenhart and W. Cyre (Eds.). Proceedings of the Seventh International Conference on Conceptual Structures (ICCS'99), July 12-15, Blacksburg, VA, USA. Lecture Notes in Artificial Intelligence, no. 1640. Berlin: Springer-Verlag, pp. 272-296, 1999.

A Hybrid Approach to Extract and Classify Relation from Biomedical Text

Abdul Wahab Muzaffar[1], Farooque Azam[1], Usman Qamar[1], Shumyla Rasheed Mir[2] and Muhammad Latif[1]

[1]Dept. of Computer Engineering, College of E&ME, National University of Sciences and Technology (NUST), H-12, Islamabad, Pakistan

[2]Dept. of Computer Science and Software Engineering, University of Hail, Hail, Saudi Arabia

Abstract - *Unstructured biomedical text is a key source of knowledge. Information extraction in biomedical is a complex task due to the high volume of data. Manual efforts produce the best results; however, it is a near impossible task for such a large amount of data. Thus, there is a need of tools and techniques in biomedical text to extract the information automatically. Biomedical text contains relationships between entities of importance for practitioner and researcher. Relation extraction is an important area in biomedical which has gained much importance. The main work is done on rule based and machine learning techniques in biomedical relation extraction. Recently the focus has changed to hybrid algorithms which have shown better results. This research proposed a hybrid rule based approach to classify relations between biomedical entities. This approach uses Noun-Verb based rules for the identification of noun and verb phrases. It then uses support vector machine, a machine learning technique to classify these relations. Our approach has been validated on standard biomedical text corpus obtained from MEDLINE 2001, an accuracy of 94.91%.*

Keywords: Relation Extraction; NLP; SVM; Classification

1 Introduction

With the huge information hidden in the biomedical field, in the form of publications, that is growing exponentially, it is not possible for researchers to keep himself updated with all developments in specific field [1, 2]. The emphasis of biomedical research is shifting from individual to whole systems, with the demand of extracting relationships between entities, e.g. protein-protein interaction, diseases-genes) from biomedical text to produce knowledge [3, 4]. Manual efforts to transform unstructured text into structured are a laborious process [5]. Automatic techniques for relation extraction are a solution to the problem. [6].

Numerous relation extraction techniques for biomedical text have been proposed [8-10]. These techniques can be categorized into four major areas i.e. co-occurrence based, pattern-based, rule-based, and ML-based approaches. Co-occurrence is the simplest technique that identifies entities co-occurs in a sentence, abstract or document [11]. Pattern-based systems use a set of patterns to extract relations; these patterns can be manually defined patterns or automatically generated patterns. Manually define patterns involve domain experts for defining patterns and is the time consuming process and have low recall [12]. To increase recall of manually generated patterns automatically generated patterns can be used. An automatic pattern generation can be used bootstrapping [13] or generate directly from corpora [14]. In rule-based systems set of rules can be built to extract relations [15, 16]. Rules-based systems can also be manually defined and automatically generated from training data. When the annotated corpora on biomedical is available machine learning based approaches become more effective and ubiquitous [17, 18].

Most approaches use supervised learning, in which relations extraction tasks are modeled as classification problems. Broadly, any relation extraction system consists of three generalized modules i.e. text preprocessing, parsing and relation extraction. In this paper, we proposed a hybrid approach to extract relations between disease and treatment from biomedical text.

2 Related Work

Research proposed by [19] focused on a hybrid approach to discover semantic relations that occur between diseases and treatments. Cure, Prevent and Side Effects are the semantic relations that have considered to be extracted between entities (Disease, Treatment). The authors claimed better results compared to previous studies done on this topic. Results show different figures for each of the three relations mentioned: Accuracy for Cure relation is 95%, Prevent relation has 75% Accuracy, 46% accuracy for Side Effect relation has been claimed.

This paper [20], primarily focused to extract the semantic relations from biomedical text. Relations are extracted between

diseases and treatments entities only. The authors propose an approach, which is a hybrid in nature, having two different techniques, to extract the semantic relations. In first technique relations extracted by pattern based on human expertise and in second one machine learning technique based on support vector machine classification. This new hybrid approach mainly relies on manual patterns when available relations examples are less, while feature values are used more when the number of available relations examples are sufficient. The authors claimed an overall F-measure of 94.07% for a cure, prevent and side effect relation.

3 Proposed Framework

The proposed work is a hybrid approach which uses rule-based and machine learning technique for the extraction of semantic relations between disease and treatment entities from biomedical text. High level view of the proposed approach is as under:

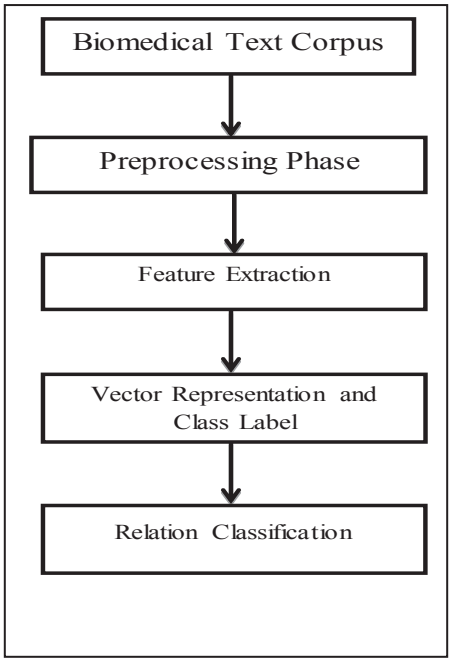


Fig. 1: Proposed Relation Extraction Approach

The framework is divided into five major steps as follows:

- Biomedical Text Corpus
- Preprocessing phase
- Feature Extraction
- Vector Representation and Class Label
- Relation Classification

Detailed flow of the proposed framework is given in Fig. 2 below:

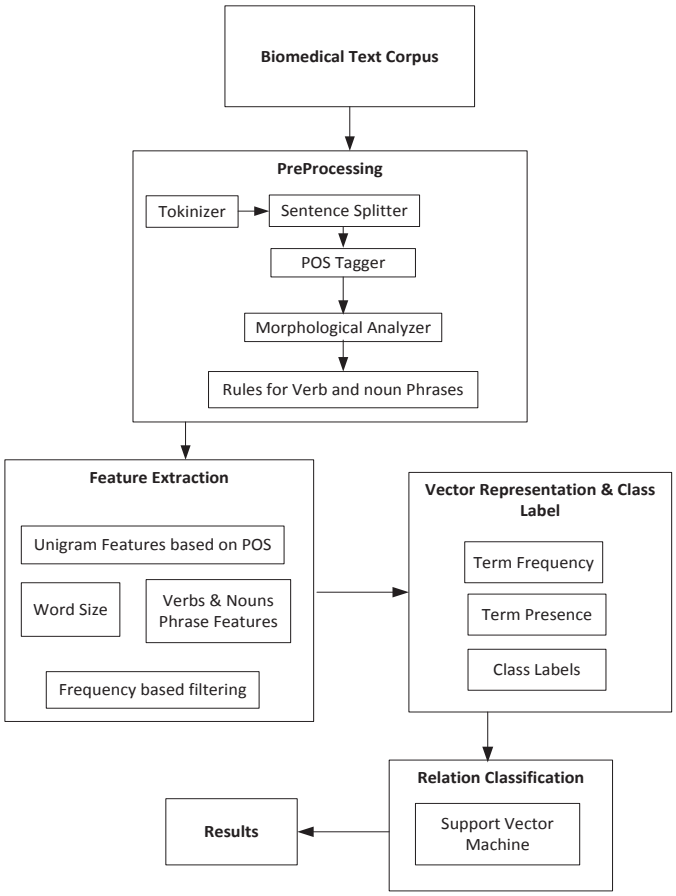


Fig. 2: Detailed Proposed Relation Extraction Framework

3.1 Biomedical Text Corpus

We used the standard text corpus that is obtained from [21]. This corpus/ data set contain eight possible types of relationships, between TREATMENT and DISEASE. This dataset was collected from Medline 2001 abstracts. Relations are annotated from sentences taken from titles and abstracts.

Table I, presents the original data set, as published in previous research showing relationships and number of sentences.

TABLE I. ORIGINAL DATASET [21]

Sr #	Relationship	No of Sentences
1	Cure/ Treat for Dis	830
2	Prevent	63
3	Side Effect	30
4	Disonly	629
5	Treatonly	169
6	Vague	37
7	To See	75
8	No Cure/ Treat No for Dis	4
9	None	1818
Total		**3655**

3.2 Preprocessing phase

In this phase following preprocessing steps has been carried out on the text corpora. We used GATE 7.1 [22, 23], for the preprocessing of corpus and used the ANNIE [23] plug-in of GATE. Following modules of ANNIE are used in this research to tag the corpus for further used of relation extraction phase:

This phase was specially aimed to extract noun and verb phrases from corpora in order to provide semantic features to classifier.

3.2.1 ANNIE English Tokenizer in GATE

The Tokenizer convert the whole text into splits the text into small tokens, i.e. different type of words, punctuations and numbers [24].

3.2.2 Annie Sentence Splitter

The sentence splitter splits the text into sentences, which is required for tagger. The sentence splitter uses a dictionary list of abbreviations to differentiate between full stops and other token types [25].

3.2.3 ANNIE POS Tagger

This module produces a part-of-speech tag and annotates each word or symbol in the text. Part of speech tags can be a verb, noun, adverb or adjectives. Tagger can be customized by changing rule set given to it.

3.2.4 GATE Morphological analyzer

The Morphological Analyzer takes as input a tokenized GATE document. It identifies the lemma and an affix of each token by considering token's part of speech tag, one at a time. These values are then added as features on the Token annotation. Morpher is based on certain regular expression rules [26]. This module uses to identify the common root of words in the text.

3.2.5 Noun-Verb-Noun Rules

We write rules and implement in JAPE [27]. These rules extract the Noun and verb phrases from text corpora. First three rules are extracting the noun phrases from text based on different criteria given in existing literature, while the last rule is extracting verb phrases. The rules are as under:

Rule: NP1

```
(
 ({Token.category == "DT"}|{Token.category == "PRP$"})?
 ({Token.category == "RB"}|{Token.category ==
"RBR"}|{Token.category == "RBS"})*
 ({Token.category == "JJ"}|{Token.category ==
"JJR"}|{Token.category == "JJS"})*
 ({Token.category == "NN"}|{Token.category == "NNS"})+
)
:nounPhrase -->
 :nounPhrase.NP = {kind="NP", rule=NP1}
```

This rule is chunking a noun phrase whenever the chunker finds an optional determiner (DT) or personal pronoun (PRP) followed by zero or more adverbs (RB,RBR,RBS) followed by zero or more adjectives (JJ,JJR,JJS) followed by one or more singular or plural noun (NN,NNS).

Rule: NP2

```
(
 ({Token.category == "NNP"})+
)
:nounPhrase -->
 :nounPhrase.NP = {kind="NP", rule=NP2}
```

This rule is simple one describing that one or more proper nouns are to be annotated as the noun phrase.

Rule: NP3

```
(
 ({Token.category == "DT"})?
 ({Token.category == "PRP"})
)
:nounPhrase -->
 :nounPhrase.NP = {kind="NP", rule=NP3}
```

This rule describes that a personal noun (PRP) can be optionally preceded by a determiner (DT).

Rule: NounandVerb

```
(
 ({NP})?
 ({Token.category == "VB"} |{Token.category ==
"VBD"}|{Token.category == "VBG"} |{Token.category ==
"VBN"} |{Token.category == "VBZ"})+
 ({NP})?
):NVP -->
 :NVP.VP = {kind="VP", rule=VP}
```

This rule chunk all the verb phrases on the basis of one or more types of verbs (VB ,VBD ,VBG ,VBN ,VBZ).

With the use of above mentioned rules we chunked all the verb phrases and noun phrases from the text. Table II demonstrates the 10 examples of both most frequent phrases occurred in the text.

TABLE II. TOP TEN VP AND NP

S/No	VP	NP
1	be treat	the treatment
2	be perform	the use
3	to evaluate	the effectiveness
4	be administer	a combination
5	to prevent	surgical management
6	to moderate	this retrospective study
7	to determine	the risk
8	should be consider	successful treatment
9	to examine	the purpose

10	be to compare	the efficacy and safety

3.3 Feature Extraction

A feature is anything that can be determined as being either present or absent in the item [34]. Feature extraction tries to find new data rows that can be used in combination to reconstruct rows of the original dataset and rather than belonging to one cluster, each row is created from a combination of the features [28].

3.3.1 Unigrams features on basis of POS

We considered unigram features on the basis of part of speech (POS). Unigram consist of a single word labeled with a single attribute. Bag-of-word features and non-conjunctive entity attribute features are unigrams [29].One vocabulary initially builds at this stage to refine unigrams features.

For unigram feature extraction, we build the vocabulary based on words of mainly four POS groups i.e. Adjectives (a), Adverbs (r), Verbs (v) and Nouns (n). Instead of each token string for a word its term root is used obtained with use of Morphological analyzer. By considering these term roots vocabulary size is much reduced. For simplicity and later comparison, we grouped the tagger POS designation into following equivalent POS.

TABLE III. TAGGER POS CONVERSION INTO SIMPLE POS

POS	Tagger's POSs
a	JJ, JJR, JJS, JJSS
r	RB, RBR, RBS
v	VB, VBD , VBG, VBN, VBP, VBZ, MD
n	NN,NNP, NNS, NNPS

3.3.2 Filtered on the basis of word size

Normally words with length less or equal to 2 are stop words or erroneous and creating noise in classification task not such as "to", "be" and "'s" etc. So we consider word size > 2 char in our unigram vocabulary.

3.3.3 Phrase Features

As shown from literature noun phrases and verb phrases are much important for relation classification we consider these phrasal features along with the unigrams.

Verbs and noun phrase chucker separate these phrases in the text based on the above described rules. We build a separate vocabulary for these phrasal features.

3.3.4 Frequency based filtering

To finally chosen features both unigrams and phrasal vocabularies are filtered by the corpus frequency with a threshold so the important terms are only selected for features.

Feature frequency > 3

Table IV shows the total no of features for unigrams, verb and noun phrases.

TABLE IV. TOTAL NO OF FEATURES IN EACH SETUP

Feature Distribution				
	Unigram	**VP**	**NP**	**Total**
Setup # 1	1267	31	101	1399
Setup # 2	673	8	37	718
Setup # 3	624	7	36	667

3.4 Vector Representation and Class Label

This phase mainly focuses on conversion of text into vectors so that it latter be used for classification. Vector is represented with the use of features and one important decision at this stage is to select each feature weight. Feature weight also affects the classification performance as shown from literature. Following are famous weight techniques used in literature:

3.4.1 Term Frequency:

How many time a feature actually occurred in the text. Easy to formulate, but suffer from limited classification performance.

3.4.2 Term Presence

Either a feature occurred in a piece of text or not, it is a binary representation. Term Presence has following advantages over the other weighting techniques.

- Easy to formulate a document or text vector
- Less processing and computation in machine learning tasks
- Better classification results as evidence from literature

3.4.3 TFIDF (Term frequency – inverse document frequency):

New weighting scheme gains importance for machine learning tasks in previous years. It is a relative weight of the feature with its document frequency. It requires more calculation to form a vector.

3.4.4 Class Labels:

Following are class labels are used with vector representation against each setup in multi class classification.

TABLE V. CLASS LABELS IN EACH SETUP

	Setups	Class Label: +1	Class Label: -1	Class Label: 0
3-Class Classification	Setup # 1	Cure	Disonly + Treatonly	Vauge
	Setup # 2	Prevent	Disonly + Treatonly	Vauge
	Setup # 3	Side Effect	Disonly + Treatonly	Vauge
	Setup # 4	Cure	Prevent	Side Effect
2-Class Classification	Setup # 5	Cure	Disonly + Treatonly	N/A
	Setup # 6	Prevent	Disonly + Treatonly	N/A
	Setup # 7	Side Effect	Disonly + Treatonly	N/A

3.5 Relation Classification

This phase mainly focuses on the classification of all those relations which exist between disease and treatment entities in the text corpora. We used SVM to classify relations in the dataset. Our main focus was to extract three main relations i.e. cure, prevent and side effect.

3.5.1 Classifier:

Relation extraction is a text classification problem, we used support vector machine (SVM) as SVMs have already been used to yield higher accuracy on related tasks like text categorization [30]. Support vector machines (SVM) have also been widely used in the PPI extraction task, and have shown competitive results over other learning methods ([31, 32]).

We use SVM classifier for our experimentation for which LIBSVM [33] is used that is integrated software for support vector classification, it supports multi-class classification. We used the LIBSVM in both setting i.e. Linear Kernel and Radial Based Function (RBF) Kernel. For linear kernel best results are obtained at c=0.5 while other parameters are on default settings, for RBF kernel best results are when g=0.05 and c=8.

4 Results and Discussion

4.1 10-fold Class Validation

The model reported here is a combination of unigram, verb and noun-phrases, biomedical, with an SVM classifier. We use cross validation when there is a limited amount of data for training and testing. We swap the role of data i.e. the data used once used for training will be used for testing and the data used for testing will use for training.

10 fold means we split the data into 10 equal partitions and the process for 10 times while ensuring that each partition is used for testing at once. Such that 10 % data for testing and 90% for training. Average the performance results of 10 x iteration to get the final results.

TABLE VI. CLASSIFICATION ACCURACY COMPARISON WITH PREVIOUS APPROACH

10 fold Cross Validation	Previous Approach Accuracy	Proposed Approach Accuracy
	Oana and Diana [20]	RBF
Cure	95%	94.91%
Prevent	75%	92.05%
Side Effect	46%	72.33%

Table VI represents a comparison of the accuracy results obtained in previous work by [20] and our proposed approach. As we can see from the table, our technique has a major improvement over previous results in setup 2 and setup 3 and a slight less improvement in setup 1. As it can be seen, our approach is very constant in all three setups. Our results improve the previous ones in setup 1, setup 2 and setup 3 with the difference varying from: accuracy in setup 1 is almost same as that of previous approach 18 percentage point improvement (setup 2) to 46 percentage point improvement (setup 3).

4.2 Discussion

Table 7 shows the comparison of previous techniques and our approach on the same data set produced by [21]. In the current settings our approach gives the best accuracy. Oana Frunza and Diana Inkpen [20], claims even better results for these relations using UMLS. Our results differ in two aspects: firstly our assumption was to extract multiple-relations-per-sentence as done by Asma Ben Abacha and Pierre Zweigenbaum [22], while the objective of the authors in [20] was the relation detection in a one-relation-per-sentence assumption. Secondly, our contribution is a hybrid approach which is combination of rule based approach and machine learning approach. Rule based techniques have higher precision but low recall while machine learning has low precision when few training examples are available. In order to increase the recall of the rule based system, machine learning approaches can be combined with rule based approaches. To increase the precision of machine learning approaches, while training set is small, we may use rule base features.

5 Conclusion and Future Work

This research focuses on hybrid framework for extracting the relations between medical entities in text documents/ corpus. We mainly investigated the extraction of semantic relations between treatments and diseases. The proposed approach

relies on (i) a rule-based technique and (ii) a supervised learning method with an SVM classifier using a set of lexical and semantic features. We experiment this approach and compared it with the previous approach [20]. The results taken by our approach shows that it considerably outperforms the previous techniques and provides a good alternative to enhance accuracy of relation extraction in the biomedical domain, if few training examples are available.

In future we are planning to test our approach with other types of relations and different corpora; we will also work on multi-stage classifier to enhance the performance of relation extraction.

6 References

[1] Jensen LJ, Saric J, Bork P: Literature mining for the biologist: from information retrieval to biological discovery. Nature Review 2006, 7:119-129.

[2] Ananiadou S, Kell DB, Tsujii J-ichi: Text mining and its potential applications in systems biology. Trends in biotechnology 2006, 24:571-9.

[3] Chapman WW, Cohen KB: Current issues in biomedical text mining and natural language processing. Journal of biomedical informatics 2009, 42:757-9.

[4] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB: Frontiers of biomedical text mining: current progress. Briefings in bioinformatics 2007, 8:358-75.

[5] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-ruxandra, Simonis N, Rual Jfrançois, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M: Literature-curated protein interaction datasets perspective. Nature Methods 2009, 6:39-46.

[6] Erhardt R a-a, Schneider R, Blaschke C: Status of text-mining techniques applied to biomedical text. Drug discovery today 2006, 11:315-25.

[7] Zhou D, He Y: Extracting interactions between proteins from the literature. Journal of Biomedical Informatics 2008, 41:393-407.

[8] Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T: All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics 2008, 9 Suppl 11:S2.

[9] Kilicoglu H, Bergler S: Adapting a General Semantic Interpretation Approach to Biological Event Extraction. In Proceedings of BioNLP Shared Task 2011 Workshop. 2011:173-182.

[10] Baumgartner WA, Cohen KB, Hunter L: An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. Journal of biomedical discovery and collaboration 2008, 3:1.

[11] Garten Y, Coulet A, Altman RB: Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics 2010, 11:1467-89.

[12] Hakenberg J: Mining Relations from the Biomedical Literature. PhD Thesis 2009:179.

[13] Wang H-C, Chen Y-H, Kao H-Y, Tsai S-J: Inference of transcriptional regulatory network by bootstrapping patterns. Bioinformatics 2011, 27:1422-8.

[14] Liu H, Komandur R, Verspoor K: From Graphs to Events : A Subgraph Matching Approach for Information Extraction from Biomedical Text. In Proceedings of BioNLP Shared Task 2011 Workshop. 2011:164-172.

[15] Jang H, Lim J, Lim J-H, Park S-J, Lee K-C, Park S-H: Finding the evidence for proteinprotein interactions from PubMed abstracts. Bioinformatics 2006, 22:e220-6.

[16] Ono T, Hishigaki H, Tanigami A, Takagi T: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2001, 17:155-61.

[17] Kim J-J, Zhang Z, Park JC, Ng S-K: BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. Bioinformatics 2006, 22:597-605.

[18] Giuliano C, Lavelli A, Romano L, Sommarive V: Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In ACL 2006. 2006, 18:401-408.

[19] Oana Frunza and Diana Inkpen, "Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences", Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 91–98, Uppsala, Sweden, 15 July 2010.

[20] Asma Ben Abacha and Pierre Zweigenbaum," A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts", CICLing 2011, Part II, LNCS 6609, pp. 139–150, 2011.c_Springer-Verlag Berlin Heidelberg 2011.

[21] Rosario B. and Marti A. Hearst. 2004. Classifying semantic relations in bioscience text. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 430.

[22] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with

GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854 — http://tinyurl.com/gate-life-sci/

[23] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311. Available from Amazon. BibTex.

[24] https://gate.ac.uk/sale/tao/splitch6.html#x9-1300006.2

[25] https://gate.ac.uk/sale/tao/splitch6.html#x9-1400006.4

[26] https://gate.ac.uk/sale/tao/splitch23.html#x28-55200023.12

[27] H. Cunningham and D. Maynard and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Technical report CS—00—10, University of Sheffield, Department of Computer Science, 2000.

[28] [28] Wu, L., Neskovic, P.: Feature extraction for EEG classification: representing electrodeoutputs as a Markov stochastic process. In: European Symposium on Artificial Neural Networks, pp. 567–572 (2007)

[29] Jing Jiang and ChengXiang Zhai," A Systematic Exploration of the Feature Space for Relation Extraction", in the proceeding of NAACL/HLT, pp. 113-120, Association of computational linguistics, Michigan (2005).

[30] Thorsten Joachims, "Text categorization with support vector machines: learning with many relevant features", in Proceedings of ECML-98, 10th European Conference on Machine Learning, Claire Nédellec and Céline Rouveirol, Eds., Heidelberg et al., 1998, pp. 137–142, Springer.

[31] Miwa M, Sætre R, Miyao Y, Tsujii J: A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2009:121-130.

[32] Kim S, Yoon J, Yang J, Park S: Walk-weighted subsequence kernels for protein protein interaction extraction. BMC bioinformatics 2010, 11:107.

[33] Chang, Chih-Chung and Lin, Chih-Jen "{LIBSVM}: A library for support vector machines", ACM Transactions on Intelligent Systems and Technology, volume = 2, issue = {3}, year = {2011}, pages = {27:1--27:27}, Available at: Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Real-Time Motion Detection Using Low-Resolution Web Cams

Mark Smith
University of Central Arkansas
Conway, Arkansas 72035

Abstract

The analysis of motion detection within video sequences has increased dramatically in recent years. One primary application of motion detection has been surveillance systems that utilize video cameras. In recent years, a new interest in surveillance systems using lower resolution webcams has increased requiring additional considerations not found in higher resolution systems. This paper examines this problem by implementing a complete surveillance system using the popular Tenvis JPT315W web camera. The motion detection algorithms are initially implemented using the standard MPEG-7 descriptors commonly used for analyzing video sequences and creating video database systems. The results of a single MPEG-7 descriptor are generally not adequate for detecting all types of motion under varying lighting conditions. This work introduces a system that intelligently combines multiple descriptors in a voting algorithm that provides more accurate results than merely using one such descriptor. An analysis on the most beneficial descriptors is presented with a ranking provided for these descriptors. Results are provided for real time videos collected from various locations undergoing a wide range of lighting conditions over a 24 hour time period. An iPhone App has also been implemented allowing access to the system remotely.

1. Introduction

Video surveillance has become one of the most important applications of motion detection and video processing systems have provided numerous applications for identifying motion [1,4]. Many popular algorithms exist using techniques similar to identifying hard-cuts (i.e., instantaneous changes) in motion picture films. These instantaneous changes result in adjacent frames of the digital video sequence undergoing significant and easily recognizable changes often detected by corresponding pixel analysis usually consisting of color differences. These algorithms have been marginal at best when applied to systems consisting of lower resolution off-the-shelf webcams undergoing varying lighting conditions. Webcams are commonly available to most consumers often making them the camera of choice due their accessibility and ease of use and integration with the overall computer system. The lower resolution of the frames produced by these webcams often introduces noise that often results in many false identification of motion [3]. This work introduces a new motion algorithm robust enough to filter the noise from the lower resolution images as well as that introduced from the effects of lighting conditions. This system was provided on MacOS and the mobile app was implemented for iOS using XCode. The system developed from this research and presented in this paper is separated into the following sections:

- MPEG-7 Overview
- Edge Histogram Features
- Gabor Filter Features
- Parametric Motion Features
- Motion Activity Features
- Extraction of Features from Individual Frames in XML format
- Classification algorithm used for Motion Detection between adjacent frames

This paper will explore each of these items and the subsequent algorithm implementation in detail.

2. MPEG-7 Overview

Using a set of reliable tools is a critical necessity in starting this project. An important first step is to integrate this MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), the committee that also developed the Emmy Award winning standards known as MPEG-1 and MPEG-2, and the standard[11,14,17]. Mpeg-1 and MPEG-2 standards made interactive video on CD-ROM and Digital Television possible. MPEG-4 provides the standardized technological elements enabling the integration of the production, distribution and content

access paradigms of the fields of digital television, interactive graphics and interactive multimedia [5].

There are many descriptors available within MPEG-7, but this paper only focuses on two general category – texture and motion. These features are selected due to their resilience to lighting and color changes that a scenes typically undergoes during different transitions of a given day. Applying a histogram equalization filter also improves the results by accounting for these lighting changes and restoring many edge details lost from the dimming of the image. The central idea for this algorithm is that texture and motion features are extracted from each frame of the video and then compared in an intelligent manner via a discriminant voting algorithm thus identifying if a given scene has significantly changed [2]. Since there is no camera motion, there is no need to first segment the video into shots or individual scenes; the webcam is providing a continuous scene for as long as the video is sampled. The texture and motion features selected from the MPEG-7 descriptors are:

- Edge Histogram
- Homogenous Texture (Gabor) Filters
- Parametric Motion
- Motion Activity

Each of these features are discussed in the sections that follow and then combined into an intelligent voting algorithm used for identifying motion between adjacent frames.

3. Edge Histogram Features

The edge histogram feature specifies the spatial distribution of the following five edge types shown in Fig. 1

(a) Vertical (b) Horizontal (c) Diagonal (d) Diagonal

Fig. 1 Edge Histogram

This feature is selected as a candidate for computing this texture measurement because of its compact size (5-dimensions) and its very efficient algorithm implementation [9]. The edge histogram is also one of the standard features used for video retrieval applications as described in the MPEG-7 standards [12,15,16]. The algorithm utilized in calculating the edge histogram features is described as follows:

a. The binary mask for a chosen object is generated. The 256-level gray-scale image is generated from the original RGB color frame. The gray-scale value for each pixel is computed as

$$gray = \frac{R+G+B}{3} \qquad (1)$$

Where R, G, and B are the color components for each pixel extracted from the original color image.

b. Next, a 4x4 mask corresponding to each edge category – vertical, horizontal, diagonal, and off-diagonal – is convolved with the object's gray-level region. This convolution process is expressed by equation (2) as

$$h_k = \sum_{i=0}^{3}\sum_{j=0}^{3} m_k g \qquad (2)$$

Where mk is one of the 4 edge masks, g is the gray-level of a pixel associated with an object, and hk is the result of the convolution. Boundary regions not fully enclosing the 4x4 mask are ignored in this computation.

c. The edge type providing the maximum value resulting from this convolution is then noted. This value is referred to as hkmax in the discussion that follows.

d. If hkmax exceeds an empirically determined threshold, the corresponding edge type is considered detected and the region is classified to the edge mask which generated the maximum convolution value hkmax. If hkmax does not exceed the threshold, the region is classified as a non-directional edge (Fig. 1(e)).

e. Steps c,d, and e are repeated for all non-overlapping 4x4 blocks of the region's interior. The four edge types and the non-directional edges are accumulated for the object resulting in a five-bin histogram.

f. Steps c,d,e, and f are then repeated for all other objects in the frame.

g. All steps (a-g) are then repeated for all frames in the sequence.

This process results in a 5-dimensional edge histogram computed for each frame . The value EHi given in (3) is based on the normalized Euclidean distance of the edge histogram computed between adjacent frames pi and ci is given as:

$$EH_i = \frac{\| eh_{pi} - eh_{ci} \|}{\| eh_{pi} \| + \| eh_{ci} \|} \quad (3)$$

Where eh_{pi} is the edge histogram computed for the previous frame, eh_{ci} is the edge histogram computed for the current frame, and EH_i is the texture measurement of frame i computed between adjacent frames.

4. Homogenous Texture (Gabor) Filters

Homogeneous texture has emerged as an important visual primitive for searching and browsing through large collections of similar looking patterns [2]. An image can be considered as a mosaic of homogeneous textures so that these texture features associated with the regions can be used to index the image data. For instance, a user browsing an aerial image database may want to identify all parking lots in the image collection. A parking lot with cars parked at regular intervals is an excellent example of a homogeneous textured pattern when viewed from a distance, such as in an Air Photo. Similarly, agricultural areas and p vegetation patches are other examples of homogeneous textures commonly found in aerial and satellite imagery. Examples of queries that could be supported in this context could include "Retrieve all Land- Satellite images of Santa Barbara which have less than 20% cloud cover" or "Find a vegetation patch that looks like this region". To support such image retrieval, an effective representation of texture is required. The Homogeneous Texture Descriptor provides a quantitative representation using 62 numbers (quantified to 8 bits each) that is useful for similarity retrieval [10]. The extraction is done as follows; the image is first filtered with a bank of orientation and scale tuned filters (modeled using Gabor functions) using Gabor filters. The first and the second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor. The number of filters used is 5x6 = 30 where 5 is the number of "scales" and 6 is the number of "directions" used in the multi-resolution decomposition using Gabor functions. An efficient implementation using projections and 1-D filtering operations exists for feature extraction. The Homogeneous Texture descriptor provides a precise quantitative description of a texture that can be used for accurate search and retrieval in this respect. The computation of this descriptor is based on filtering using scale and orientation selective kernels. A diagram illustrating the implementation of the Gabor Filter is shown below:

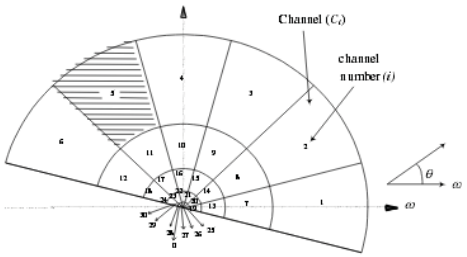


Fig. 2 Gabor Filters

allow the system to filter outliers or incorrect incidents from those newly occurring events.

5. Parametric Motion and Motion Activity

Parametric motion models have been extensively used within various related image processing and analysis areas, including motion-based segmentation and estimation, global motion estimation, mosaicking and object tracking [6]. Parametric motion models have been already used in MPEG-4, for global motion estimation and compensation and sprite generation. Within the MPEG-7 framework, motion is a highly relevant feature, related to the spatial-temporal structure of a video and concerning several MPEG-7 specific applications, such as storage and retrieval of video databases and hyperlinking purposes. Motion is also a crucial feature for some domain specific applications that have already been considered within the MPEG-7 framework, such as sign language indexation [13]. The basic underlying principle consists of describing the motion of objects in video sequences as a 2D parametric model.

A human watching a video or animation sequence perceives it as being a slow sequence, fast paced sequence, action sequence etc. The activity descriptor captures this intuitive notion of 'intensity of action' or 'pace of action' in a video segment [7]. Examples of high 'activity' include scenes such as 'goal scoring in a soccer match', 'scoring in a basketball game', 'a high speed car chase' etc.. On the other hand scenes such as 'news reader shot', 'an interview scene', 'a still shot' etc. are perceived as low action shots. Video content in general spans the gamut from high to low activity, therefore we need a descriptor that enables us to accurately express the activity of a given video sequence/shot and comprehensively covers the aforementioned gamut [8]. The activity descriptor is useful for applications such as video re-purposing, surveillance and fast browsing, An example of motion activity computed for a scene is shown in Fig. 3:

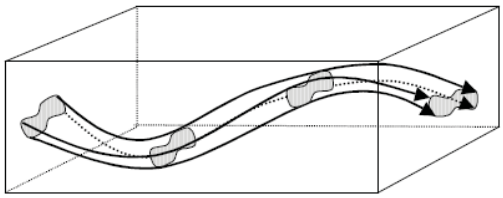


Fig. 3 Motion Activity

6. Extraction of Features

The 4 MPEG-7 features are extracted in the form of vectors and stored in a file formatted using the XML Language. The vectors generated for each of the MPEG-7 features are shown below:

- 5 features for Edge Histogram
- 62 features for Homogenous Texture
- 7 features for Parametric Motion
- 5 features for Motion Activity

The vectors are formatted within an XML file thus allowing for ease of organization and retrieval. An example of the XML extracted for the above features are shown below:

```
<?xml version="1.0"?>
<frame name="142819" date="4-28-2015 14:17:21">
  <feature name="EH"/>232 431 1234 891 22103
   </feature>
  <feature name="HT"/>6 21 4 32……(others)
   <//feature>
  <feature name="PM"/>1.5 2.3 8.2 ….(others)
   </feature>
  <feature name="MA"/>0 4.7 9.3 12.4 (others)
   </feature>
 </frame>>
 …..(others follow)
```

The features are utilized in an algorithm consisting of a series of equations used in identifying the motion computed between adjacent frames. The next section describes the algorithms used for classifying the motion between adjacent frames.

7. Motion Identification and Classification

. The motion and texture features are extracted for each frame as specified in [2,3] are grouped into corresponding vectors given as equation (4) given below . The corresponding vector distances are computed between adjacent frames (*I* and *j*) for each feature as shown below*:*

$$V_{ij} = \frac{\sum_{k=0}^{N-1} f_{ik} - f_{jk}}{\| f_i \| + \| f_j \|} \quad (4)$$

where *i* and *j* are the adjacent frames, and *k* is the *kih* feature and then compared by computing the Normalized Euclidean difference between each set. The mean for each vector set is next computed over each frame as shown below in 5:

$$\mu_i = \frac{\sum_{k=0}^{N} V_{ik}}{N} \quad (5)$$

The mean for each vector set is updated for each frame that is encountered. The standard deviation between all frames for a given vector is then computed as shown below:

$$\sigma_i = \frac{\sqrt{(V_i - \mu_i)^2}}{N-1} \quad (6)$$

The adaptive threshold between frames is used when classifying the motion between adjacent frames. Motion is assigned to the second frame if the vector difference is greater than the adaptive threshold given in (7) as:

$$V_{ij} > 2\sigma_i + \mu_i \quad (7)$$

Equations (4) thru (7) are repeated for each of the 4 vector sets – Edge Histogram, Homogenous Text, Parametric Motion, and Motion Activity. Equation (7) is computed for all vector sets and for all cases that (7) is true, is maintained If 3 out of 4 of the vector sets are true, motion will be assigned to the second frame. This result can be expressed in the following equation as:

$$motion_{p+1} = \sum_{i=0}^{N} T_{c_{ip}} \bigcap T_{c_{ip+1}} > 3 \quad (8)$$

The major assumption for this algorithm is that a series of frames with no motion are encountered first before any motion occurs. The series of non-motion frames are used for initializing the system and setting a baseline and therefore providing training to this system, so when motion is encountered, it is easily identified and classified with minimum errors.

8. Testing and Results

This system was initially tested on a series of standard MPEG videos containing a series of frames undergoing motion. Table 1 shown below illustrates

the results of the system The 2nd column contains the total number of frames processed, the 3rd column indicates the number of frames correctly classified as either motion/no motion, while the 4th column indicates the number of frames incorrectly classified as either motion/no motion. The Percent Correct is given as the right most column.

Table 1 Results

Video	Total	Correct	False	Percent Correct
Happy Granny	62	59	3	95.1%
Foreman	43	42	1	97..6%
MotorCycle	39	35	4	89.7.%
News	137	130	7	94.9%

Most of the errors occur at the outliers – either when the motion occurs at the beginning or the end of the video clip. Also very slight motion is identified and classified as motion.

The system was also tested with a low resolution Tenvis JPT315W web cam producing an image with a 160x120 resolution. The web cam was mounted in the author's office and produced/transmitted images every 12 seconds to a web server running this system for analysis. The test was performed over a 24 hour period where the web cam underwent a variety of lighting conditions with over 7000 frames transferred and tested. The results of the web cam tests were similar to that of the standard video with a very high percentage of the video clip classified correctly as either having motion or not having motion.

The results appear very promising illustrating the accuracy of this system. The error rate is well within bounds and provides users with a very accurate motion detection system for low-resolution web cams.

9. References

[1] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 6, pp. 939-954, 2001.

[2] Air Pressure: Why IT Must Sort Out App Mobilization Challenges". InformationWeek. 5 December 2009.

[3] E. D. Gelasca, E. Salvador and T. Ebrahimi, "Intuitive strategy for parameter setting in video segmentation," Proc. IEEE Workshop on Video Analysis, pp.221-225, 2000.

[4] MPEG-4 , "Testing and evaluation procedures document", ISO/TEC JTC1/SC29/WG11, N999, (July 1995).

[5] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences," *ICASSP '97 Proceedings*, pp. 2657 – 2660, 1997.

[6] T. Aach, A Kaup, and R. Mester, "Statistical model-based change detection in moving video," *IEEE Trans. on Signal Processing*, vol. 31, no 2, pp. 165-180, March 1993.

[7] L. Chiariglione-Convenor, technical specification *MPEG-1 ISO/IEC JTC1/SC29/WG11 NMPEG 96*, pp. 34-82, June, 1996.

[8] MPEG-7, ISO/IEC JTC1/SC29/WG211, N2207, Context and objectives, (March 1998).

[9] P. Deitel ,*iPhone Programming*, Prentice Hall, pp. 190-194, 2009.

[10] C. Zhan, X. Duan, S. Xu., Z. Song, M. Luo, "An Improved Moving Object Detection Algorithm Based on Frame Difference and Edge Detection," 4th International Conference on Image and Graphics (ICIG), 2007.

[11] R. Cucchiara, C. Grana, M. Piccardi, Member and A. Prati, "Detecting Moving Objects, Ghosts, and Shadows in Video Streams," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1337-1342, October, 2003.

[12] F. Rothganger, S. Lazebnik, C. Schmid and J. Ponce, "Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no.3, pp. 477-491, March 2007.

[13] Neil Day, Jose M. Martinez, "Introduction to MPEG-7", ISO/IEC/SC29/WG11 N4325, July, 2001.

[14] M. Ghanbari, Video Coding an Introduction to standard codecs, Institution of Electrical Engineers (IEE), 1999, pp. 87- 116.

[15] L. Davis, "An Empirical Evaluation of Generalized Cooccurrence Matrices," IEEE Trans. Pattern Analysis and Machine Intelligence, vol 2, pp. 214-221, 1981.

[16] R. Gonzalez, Digital Image Processing, Prentice Hall, 2nd edition, pp. 326-327, 2002

[17] K. Castelman,Digital Image Processing, Prentice Hall, pp. 452-454, 1996.

Market Research of the Innovative Smelling Smartphone

Raed Al-Wishah
Princess Sumaya University for Technology, Amman, Jordan 11941
r.wishah@psut.edu.jo

Abstract

The Smartphone has become a big part of everyone's lives; they are relied on for just about everything. With the increase in technology, smart phones have obtained the capability to allow the user to do just about anything. There is a wide range of different Smartphone applications to choose from. Smartphone's have gained the capability of using biometrics for several applications. Many different detectors have been developed over the past couple of years, these detectors range from fingerprint sensors, iris recognition and now odor detecting sensors. This research is a study on how to market a smart phone device that detects smells; the following question explains the idea *"Imagine your Smartphone could detect odors/smells, like a sniffing dog. How would you use this function in your everyday life?"* Most people would want to use the application to check for food freshness, alcohol, smoke, bad breath and allergies.

Keywords: smart phones, Socially Ethical, Smartphone technology.

1 Introduction

The electronic nose is a new technology in a form of a device that helps identify the components of an odor. The idea of the electronic nose has been around for the past couple of years but has recently been implemented into usable form. Many different world wide organizations have done researches to identify whether or not this technology could be implemented into usable portable forms [1]. Discussions have been made about whether or not this technology should be used in either internal or external form. Based on studies, some of the final consumers prefer for this technology to be used in the form of an internal device where it is already built into their smart phone, and others prefer for it to be attached as an external device [2].

Odors are composed of molecules of different sizes and shapes. Each and every single one of these molecules is detected differently by the human nose. After the receptor receives the molecule it sends a signal to the human brain, where it is then identified. The electronic nose is a technology based on the idea of biometrics, which is the study of measurable biological characteristics. It also involves human-made applications patterned on natural phenomena's.

Ever since 1970, studies have been made to figure out whether or not the idea of the electronic nose could be implemented in a user friendly way [3]. Agencies/Companies/Organizations such as NASA, Apple, Samsung, blackberry and many others have done different studies on the implementation of the electronic nose into smart phones. These studies are still on-going till today, even though some have already been implemented; these different companies are still searching for ways to develop this technology more and more. In such ways as, detecting more odors according to customer preferences [3].

This innovative technology has many strengths and weaknesses. These weaknesses include loss of sensitivity or high concentrations, inability to detect all odors, relatively short life, having high sensitivity that is much higher than that of the human nose [4]. Other weaknesses or limitations include some customers not accepting the idea, the accuracy of the detector and pricing. On the other hand this technology has many optimistic viewpoints. Such as medically to alert allergens of certain odors that they are allergic to, detecting out-dated food, detecting bad odors, and other detections. Other strengths of this technology include its user friendly and easy to use, mostly as a form of protection and as a form of entertainment [4].

The core idea of our research is that a smart phone must have two components; the first component is hardware as a device whether internal or external. The other component is software in the form of a smart phone application.

Based on our research the odors that most customers would like to be detected by their smart phones are dangers, gasses, smoke, body smell, alcohol, bad breath, car pollution and body odor etc [5].

The rest of the paper is organized as follows: Section 2 describes the history of the Smartphone background and literature. Section 3 presents the research design. Section 4 describes the result analysis and discussions Section 5 concludes the work presented in this paper.

2 The History of the Smartphone:

In the early hours, devices that were integrated into the telephone were primary envisioned back in 1973. Such technology was proposed for commercial transactions in the early 1990s. The name "Smartphone" first arrived in 1997, and that was when Ericsson has depicted its GS 88 concept as a Smartphone [6-8].

Forerunners came right after when Ericsson introduced their concept of the Smartphone, the first mobile phone. The first mobile phone to merge the characteristics and features of the PDA was the IBM original model that was launched in 1992 [9]. Also receiving the approval that year at the COMDEX computer industry trade fair. A revised adaption of such a product was commercially launched and marketed to final consumers back in 1994-1995 by BellSouth, holding the main term of the Simon Personal Communicator [10]. This Communicator was the first device that may be fittingly recognized as a "Smartphone". Although that name was not marked until later. Moreover, its capability to do and accept cellular phone incoming calls, this model of Simon had the ability to generate and receive e-mails and faxes via the screen display touch [10-12].

At the end of the 1990s, several mobile phone users started to have a separate devoted PDA device. Earlier types of the operating systems like BlackBerry, windows and others. Later on in 1996; Nokia had produced the great leading phone the NOKIA 9000.Based on its sales levels it was a big success. It indeed was a palmtop computer-styled phone, intergraded with a PDA device from a well-known company in the field of technology that is HP. Earlier on original models, only a couple of devices were fit together by a pivot. During earlier prototypes, the two devices were fixed together via a pivot and by that it turned out to be acknowledged as a clamshell design. At that time the initial design started with opening the display on the top surface with a tangible keyboard on the bottom. Email and text based web surfing were set by GEOS operating system. Getting into the early 2000's, the Ericson R380 was new in the market by the Ericson Mobile Communication and it was also the first leading device that was commercially successful as a Smartphone. It included the several operations of a mobile phone and also a personal digital help. PDA, helped restricted operations of such a mobile phone like web surfing with an opposed touch screen. Later on in 2001, Palm incorporation's had released the Kyocera that was integrated with a PDA device in the mobile phone and operated by Verizon. It had also taken web surfing to the next level. Smartphone's even before the Android operating system, BlackBerry and the IOS, generally worked on Symbian that had been initially launched by Psion and it was one of the most broadly used smart phone operating systems worldwide till the last quarter of 2010 [13].

After the year 2006, new corporations had started to show up in a very strong manner. One of the leading companies that arose later on is Apple Inc. The iphone had taken a huge place in the market and was also very attractive. It was also the first Smartphone that was useable with a multi-touch interface. It was also popular for its large touch screen device, as its major interaction. The two operating systems (IOS and Windows phone) were on top of the market and had a huge demand. Then Blackberry 10 also had penetrated the market as a third player following the two main operating companies [13].

The first "Socially Ethical" smart phone was out in the market by a company called Fair-phone, at the London design Festival. Who in which made more

concerns about materials sourcing in the manufacturing industry. At the end of 2013, QS Alpha began producing a Smartphone that was designed completely for security purposes, such as identity safety. Also in 2013 the Samsung Galaxy had been released to the retail market which was a major success and the flexibility was high. Moreover, Foldable OLED Smartphone's might be a decade back due to the cost of its production**.** Relatively speaking, there is a high failure rate when making such screens. The clarity in the design, for instance the thin layer of crystal glass can be added to small screens such as watches and smart phones in which allows them to work through solar power. There is a possibility that Smartphone's might achieve a 16% higher battery life. The estimation for such a smart phone using this type of technology is expected to be in the market by 2015. Also another progressing technology concerning screens is one that allows the screen to operate through receiving Wi-Fi signals. Such a smart phone is expected to cost between 2$ to 3$ which means it's much cheaper than any other technology that is already available. During the beginning of 2014, smart phones started to be used as Quad HD, a screen that was a huge development over Apple's retina display. In fact, Quad HD is used in new TV's and computer screens for a bigger and better display. Moreover at this year, the wireless will carry on to place the core network of a Smartphone. In February 2014, some smart phones were designed to be dustproof and waterproof. According to the water-resistant fabric feature, the Sony Xperia Z1S and the Samsung Galaxy S5 have been capable getting up to one-meter of underwater depth, not to mention that there are no harms left to the phone [14].

Summarizing this useful study by the following:

- The capacity of this technology and it has a good capability
- How the new generation of micro bridge resonators is.
- The Process of making such an innovative tech.
- The measurements methods have been used.

Thus, All of that demonstrates to us that such a technology is out there and on peoples mind but not on the Smartphone so applicable connotations – The current technology prepares a good utilization of such a based micromachining methods and spreads out commercial inkjet printing for such a function of the individual detection elements. This enhances its potential adaptation by industry. The innovative concept paves the way for autonomous electronic nose systems.

According to the previous studies; it shows that such an innovative technology is already has been researched and launched however, few have done it on the Smartphone so it makes a Smartphone could detect smells and so on, what we have seen before approves that there is a possibility to make such a smelling Smartphone by integrating a connected device to it as well as the e-nose whether the device is internal or external with the phone. Our smelling Smartphone market research is to discover that it can be done and used for the final user of the Smartphone.

3 Research Design:

This study is based on both quantitative and qualitative data collection. "Quantitative data is defined as a type of information that can be counted or expressed numerically. This type of data is often collected in experiments, manipulated and statistically analyzed. Quantitative data can be represented visually in graphs, histograms, tables, and charts." Were as Qualitative data is defined as "data that can be arranged into categories that are not numerical. These categories can be physical traits, gender, colors or anything that does not have a number associated to it. Qualitative data is sometimes referred to as categorical data. Qualitative data is contrasted with quantitative data. Quantitative data sets have numbers associated with them." An example to the use of the above data is questionnaires and focus-closed groups [15].

In regular internal meetings, the methods and tools of market research were selected to support our work, depending on what the clientele bias will require. In

order to develop a starting point and to become a sense of the topic, we conducted a team brainstorming session. During the brain storming sessions; more ideas were presented and opposed, despite the fact that many of these ideas were unrealistic still they lit the light for the realistic scenarios [16].

In order to gain further knowledge we build up a focus group. During these focus group sessions we organized interviews and discussions to win new approaches, due to the fact that interest is shown from companies and clients. This also took the ideas and details to a total new level with the motivation that the team had gained from the previous held steps and acceptance.

In addition, international experts from Smartphone manufacturers, mobile operators and retailers were contacted to obtain information about licensing and know the attractiveness of the device, but unfortunately none of them answered our questions. Moreover students and many other potential users were used in this study population in order to maximize the flow of information.

4 Result Analysis and Discussions

In this market research, we have used the mix method where it consists of the qualitative and quantitative method. In the qualitative method; we have conducted a focus group that has an interviewer and a few people answering questions about the smelling Smartphone technology. In the quantitative, we have spread an online survey to different kind of people in several countries; in the survey, we have asked important questions about the Smartphone in general and if such an innovative technology is interesting

The respondents of the surveys have many characteristics. 47% of the respondents are males and 53% are females. As for the country of residence; 57% of the respondents are from Germany, 24% are from China, 6% are from Korea, 4% are from Jordan, 4% are from Bulgaria, 2% are from France and the remaining 3% are from other countries such as USA, UK, Switzerland, Canada, Palestine, Croatia, Indonesia, Liechtenstein, Morocco, Spain and Singapore. 57.3% of the respondents are in between the ages of 19 and 24. 15.8% of the respondents are between the ages of 25 and 29. 12.2% of the respondents are 40 years of age and older. 8.9% of the respondents are under 19 years of age. 3.5% are in-between 30 to 34 years old and 1.4% are between the ages of 35 to 40. As for the occupations of the respondents; 68% are students. 10% of the respondents provided no answer for their occupation. 5% claim to be “other”, 4% are employees at a university either in the academic field or administrative. 3% are employees, 2% work in the I.T sector, 2% are engineers, and the remaining 1% consists of several other occupations.

1.1. Interview Results:

The qualitative is one of our methods that we followed thus we conducted a focus group with different people and the interviewer asked six main questions. The Questions and the answers are:

1. **What do you think about the Smelling Smartphone?**

Person 1 said: "I think it's not useful. I mean it's too childish"

Person Two said: "No, I think it is really useful. Like the one app to find out if a food is fresh or not."

Person Three said: "Sure why not. Our technology has been improved a lot in only a couple of years. I think the smelling phone is a new technology with potential for innovative ideas."

2. **Do you have any new ideas on applications?**

Person 1 said: "Sound good to me. Mothers who are worried about their children would love that. You could also try to detect dangerous material in toys and clothes. Yes, I would buy that to make sure the stuff I buy for my children are not dangerous"

Person 2 said: “You mean like a pollen detector. Yes would be useful if you don’t know if it is safe to go outside today"

Person 3 said: “I think it would be nice if the smelling phone could be use like a tester for perfume."

3. **Which application is the most interesting?**

Person 1 said: "I think the food freshness app is the most interesting one."

Person 2 said: " Yes that a good idea and to detect if the food is really Bio or not."
Person 3 said: "Food Freshness and Safety, Exactly, I mean there are some people who really care about that. They would appreciate that"

4. **How much would you pay for a smelling app?**

Person 1 said: "Well, depend on the apps. For fun apps I would pay nothing."
Person 2 said: "I usually don't buy apps but I agree with Ahmad I think the apps should be an additional service for the customer."
Person 3 said: "If the app is really good and I would really use it every day like the pollen detector or the food freshness app. I would even pay 10 JOD, 10 JOD is a lot for just an app. Of course I would pay for apps if they are useful."

5. **Do you prefer the smelling device attached to your Smartphone?**

Person 1 said: "it would be better to have it inside the phone"
Person 2 said: "Yes, inside is much better"
Person 3 said: " Agree, I think everyone want the device not to be separate"

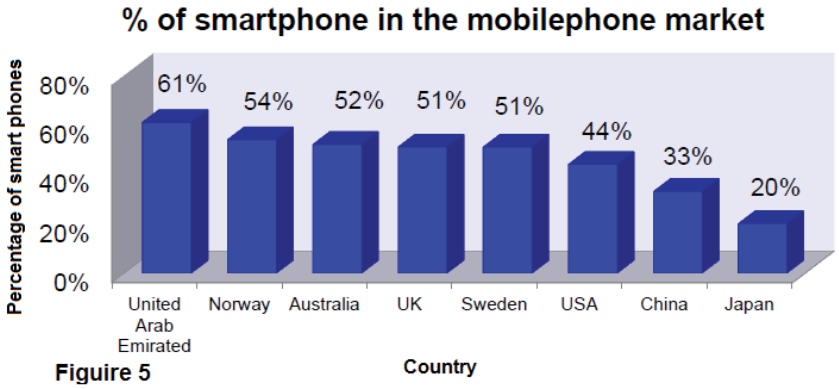
6. **Would you buy a smelling Smartphone?**

Person 1 said: "Well, a smelling would come in handy in some situation. I like the food freshness app and I would probably use it every day"
Person 2 said: "Maybe later when the technologies become standardized."
Person 3 said: "Yes, I would, could be protection and fun at the same time"
Summarizing all the questions and answers by the following:

• Some people said it is not useful childish and others said vice versa, very useful especially the food and safety apps.
• Most of the interesting apps were about the Safety and Food Freshness Apps.
• Most people are willing to pay around 7 to 10 dollars for a smelling app.
• In fact, most people prefer the device to be internal however; some said that an external device is more suitable for them
• 90% of the them said "YES, We would buy a smelling Smartphone".

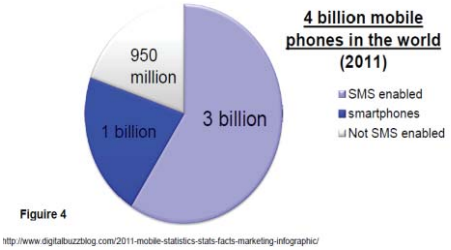
1.2. Discussions:

In the discussions part; we will be mainly discussing the survey results especially the questions of the applications; the payment and attractiveness wise, in addition we will start briefly by talking about the Smartphone market and its usages.



Figuire 5

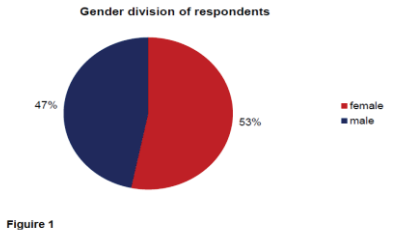
The Smartphone market has enjoyed a rapid growth over the years. For example in 2012/2013, 29% of the mobile phones in Germany are Smartphone. The use in other countries is even higher such as the UAE.



Figuire 4
http://www.digitalbuzzblog.com/2011-mobile-statistics-stats-facts-marketing-infographic/

In fact, there is a high potential for Smart phones to become the biggest part of the

mobile phone market. Talking about the Smartphone manufacturers



Figuire 1

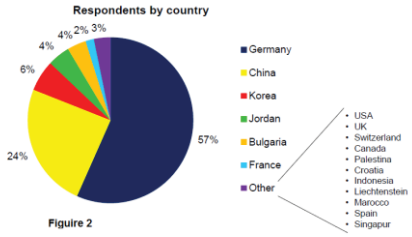
Question asked: What is your gender?

After conducting the online survey which was sent to different kind of people in different countries mainly in Jordan, Germany and China, in the survey; the team formed several important questions starting with

"What is your gender?"

The answers indeed were shocking that 47% of the respondents were males and the 53% were females.

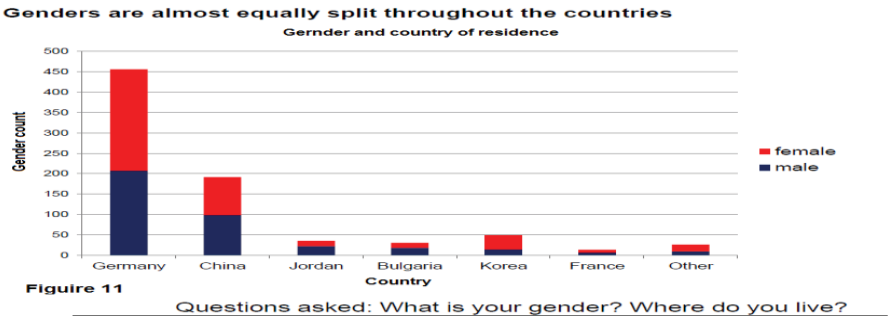
and Operating systems as well, there are seven top Smartphone manufacturers and five main operating systems on the market.



Figuire 2

Question asked: Where do you live?

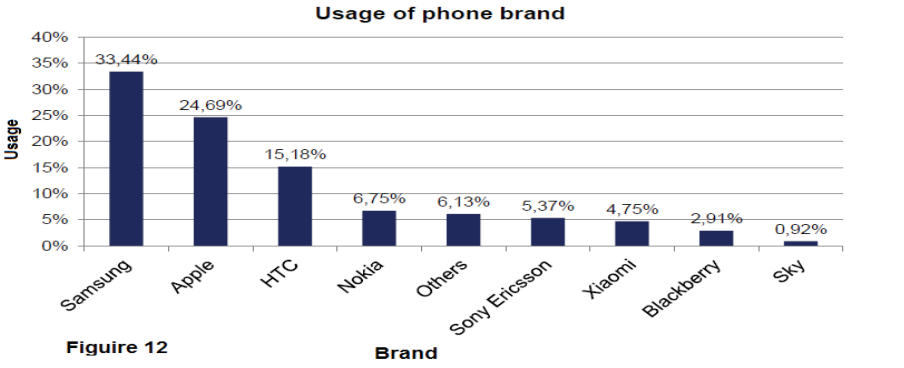
For instance, an important question was asked which is **"Where do you live?"** : So the results of the survey had fruitful answers from different countries however; most of the respondents come from Germany, China and Jordan.



Figuire 11

Questions asked: What is your gender? Where do you live?

Genders are almost equally split throughout the countries. The Smartphone usage among the respondents were more than 80% of them use a Smartphone.



Figuire 12

Question asked: Which smartphone brand do you use?

"Which Smartphone brand do you use?" the answers were as expected however mostly used Smart phones are Samsung, Apple, HTC and Nokia, the rest were minors such as Blackberry and others.

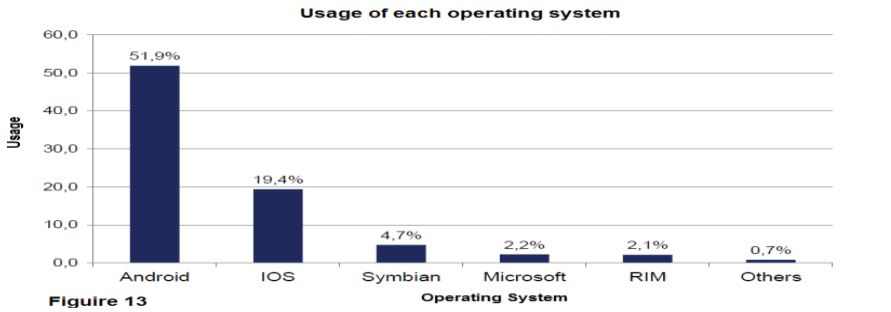


Figure 13

Question asked: Which smartphone system do you use?

The following question was about the operating system that is "**Which Smartphone system do you use?**" expectations came true about the most used ones that are the Android and IOS since most people use Samsung and Apple as well.

"**How many hours do you use your Smartphone a day?**" the answers were 68% of the respondents use their phone between 1-5 hours per day.

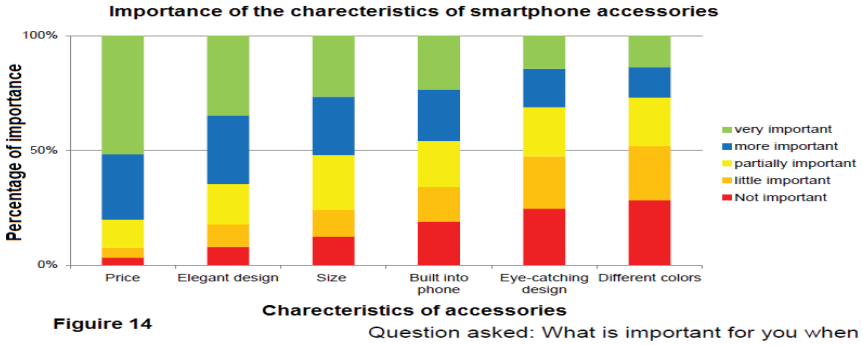


Figure 14

Question asked: What is important for you when buying electronic accessories for your smartphone?

Upon another important question about the phone accessories since this Smelling Smartphone could be held as an internal or external device, the question was "**What is important for you when buying electronic accessories for your Smartphone?**" the respondents answers were that the price is the most important trait for Smartphone accessories.

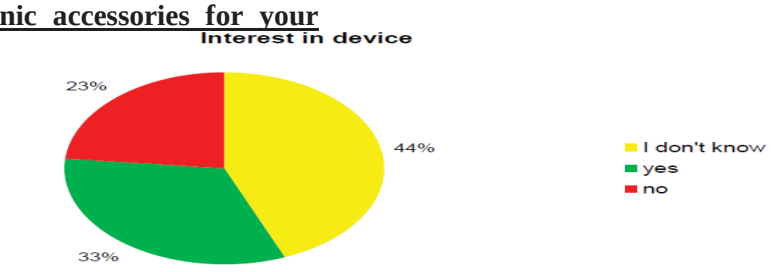
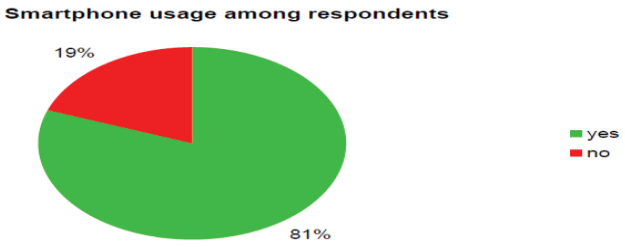


Figure 8

Question asked: Imagine your smartphone could detect smells. Would you use such a function?

The next question was "**Imagine your Smartphone could detect smells would you use such a function?**" 33% of the respondents would buy the device. And the largest group of respondents that are able to use a smelling device was from China! Demographically, there is only a slight difference between the interest of men and women. Most importantly of this question is that over 35 year olds are least interested in the device.



Figuire 6

Question asked: Do you use a smartphone?

Apps payment in general, more than 28% of the respondents are willing to pay for an app**.** Another fact is that 53% of the respondents usually pay 1€ for an app.

Finally, Here are the results of two important questions for each application in the survey; we focused on two important components that are:

1) The Attractiveness of the application
2) The Payment of the application.

The main question for the attractiveness was:**How attractive is this app for you personally?** The other main question for the payment was: **What price do you think is acceptable, after users had the chance to test the product?**

Applications results after spreading the survey:

After conducting the brainstorming sessions; we have filtered the best applications that are feasible, interesting and most importantly MobiSense has agreed on them, after that, we have done the survey and the focus group and here are the results of each application: We focused on two important components that are:

1. The Attractiveness of the application
2. The Payment of the application. The main question for the attractiveness was: **How attractive is this app for you personally?** The other main question for the payment was: **What price do you think is acceptable, after users had the chance to test the product?**

Application Results:

1) Counterfeiting App:

The attractiveness of the *Counterfeiting app* is very and more attractive for people in the age groups younger than 19 and Between 25 to 34. About the price, most people between the ages under 19 and older 40 are not willing to pay for the *Counterfeiting app.*

2) Food Freshness App:

The attractiveness of The *Food Freshness app* is most attractive to respondents under 19 however there is a high attractiveness among the other ages as well.About the price, most respondents from Jordan would not pay for the *Food Freshness app*, but ¼ of the people from Bulgaria are willing to pay more than 5 €.

3) Coffee Freshness App:

The attractiveness of the *Coffee Freshness app* is very up to more attractive for Jordan. About the price, 55% of women would pay 2-5€ and 55% of men would pay more than 5€ for the *Coffee Freshness app*.

4) Coffee and Tea Maker App:

The attractiveness of the *Coffee and Tea Maker app* has nearly the same attractiveness for Jordan as it has for France.About the price, most participants over 35 don´t think that any price is appropriate for the *Coffee and Tea Maker app*.

Car Pollution App:

The attractiveness of the car pollution app is that most people in the age group between 30-34 think that the *In Car Pollution app* is very attractive

About the price, There are more

women who would like to pay up to 2€ for the *In Car Pollution app*.

5) ***Air App***
The attractiveness of the air application is that around 70% of the Koreans think that the *Air app* is very up to more attractive .About the price, more than 50% of the Jordan people wouldn´t like to pay for the *Air app*.

6) ***Drink App***
The attractiveness of the *Drink app* is very up to most attractive to people in Jordan, Korea and China. About the price, the participants over 40 don´t think that any price is appropriate for the *Drink app*.

7) ***Smell Quiz App:***
The unattractiveness of the *Smell Quiz app* is outruns the attractiveness in all countries. About the price, more men would pay more than 5€ for the *Smell Quiz app*.

8) **Smoker App:**
The attractiveness of the *Drink app* of the *Smoker app* is very up to most attractive for people in Jordan and France.About the price, there are more women who would pay more than 5€ for the *Smoker app*.

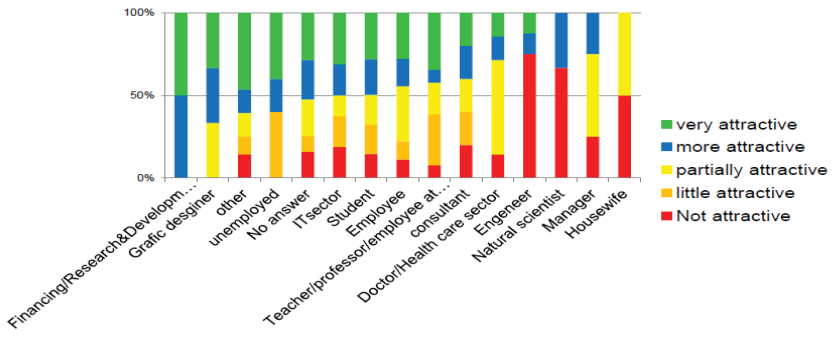
9) ***Body Smell App***:
The attractiveness of the *Body Smell app* is very up to more attractive for most of the respondents from Jordan and Korea.About the price, more than 50% of people over 40 wouldn´t pay for the *Body Smell App*.

10) ***Bad Breath App:***
The attractiveness of the *Bad Breath app* is very up to more attractive for people in the age group under 19 and Between 30 to 34. About the Price, more than half of the women would pay between 2 to 5€ for the *Bad Breath app*.

11) ***Alcohol app***:
The attractiveness of the Alcohol App is that more than half of men find the *Alcohol app* very attractive, but in general the app is also attractive to women. About the Price, There are more men who would pay 2-5€ for the *Alcohol app*.

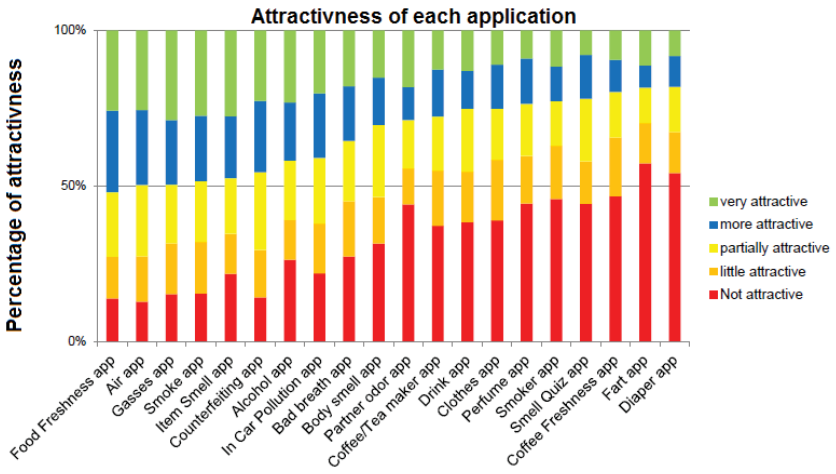
12) ***Clothes app:***
The attractiveness of the clothes app is that with an increasing age the *Clothes app* becomes more unattractive for the majority of all respondents. About the price, Hardly anyone is willing to pay more than 1€ for the *Clothes app*.

13) ***Gases App:***



Question asked: How attractive is this app for you personally?

This was indeed an important question and the team asked people working in different fields as well as doctors, students, Consultants, Managers, Housewives, Graphic designer, etc. And the results were that the *Gasses app* is very up to more attractive for all Financing/Research & Development employees.
About the price, more than 50% of men are not willing to pay for the *Gasses app*.



Figuire 18

Question asked: How attractive is each app for you?

Results approves that the Food Freshness app, Air app and Gasses apps are the three most attractive apps and more than 20% of the respondents would pay 2€ and more for a danger detection app.

5 Conclusion

Based on the high acceptance that we have received over the idea of the innovative smelling smart phone, we conclude that the device would be a success. We conclude that the device would mainly be used for the detection of danger and for general curiosity. One of the key factors that enhance the level of acceptance is that the idea of the device is very new to the common public, even though the idea has been around for years. But based on the feedback that we have received, it is concluded that in order for the device to be a success, the price of the device must be amongst a range that is affordable to the common public. Also the device should be in the form of an internal device rather than an external device. Alongside those factors it is also concluded that the device would be popular world-wide. Respondents were generally happy with all of the attributes measured in the survey and noted that the survey allowed them to think deeper into what their desires were of the devices capabilities. The one thing that they were not very pleased with was that not all applications on the device were feasible, such as for example the "perfume application".

References

[1] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International journal of human-computer studies*, vol. 43, pp. 907-928, 1995.

[2] M. N. Boulos, *et al.*, "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX," *Biomedical engineering online*, vol. 10, p. 24, 2011.

[3] F. Röck, *et al.*, "Electronic nose: current status and future trends," *Chemical reviews*, vol. 108, pp. 705-725, 2008.

[4] W. J. Harper, "The strengths and weaknesses of the electronic nose," in *Headspace analysis of foods and flavors*, ed: Springer, 2001, pp. 59-71.

[5] H. Verkasalo, *et al.*, "Analysis of users and non-users of smartphone applications," *Telematics and Informatics*, vol. 27, pp. 242-255, 2010.

[6] S. Heo, *et al.*, "Lifelog Collection Using a Smartphone for Medical History Form," in *IT Convergence and Services*, ed: Springer, 2011, pp. 575-581.

[7] J. Lee, *et al.*, "A GIS-based design for a smartphone disaster information service application,"

in *Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on*, 2011, pp. 338-341.

[8] D. Santamarta*, et al.*, "The natural history of arachnoid cysts: endoscopic and cine-mode MRI evidence of a slit-valve mechanism," *min-Minimally Invasive Neurosurgery,* vol. 38, pp. 133-137, 1995.

[9] J. Stöhr and M. Samant, "Liquid crystal alignment by rubbed polymer surfaces: a microscopic bond orientation model," *Journal of electron spectroscopy and related phenomena,* vol. 98, pp. 189-207, 1999.

[10] O. Palm, "This article's tone or style may not reflect the encyclopedic tone used on Wikipedia. See Wikipedia's guide to writing better articles for suggestions.(June 2012) A smartphone is a mobile phone built on a mobile computing platform, with more advanced computing ability and connectivity than a feature phone. The first smartphones mainly combined the functions of a personal digital assistant (PDA) and a mobile phone or camera phone. Today's models also serve to combine the functions of portable media players, low-end compact digital cameras, pocket video cameras, and GPS navigation units."

[11] I. Krajci and D. Cummings, "The Mobile Device and Operating System Landscape," in *Android on x86*, ed: Springer, 2013, pp. 9-15.

[12] P. Aswini and C. A. Kumar, "Traditional Data Delivery Scheme Using Store-Carry-Forward Protocols in Smart Phones."

[13] T. Farley, "Mobile telephone history," *Telektronikk,* vol. 101, p. 22, 2005.

[14] A. Carroll and A. Buchholtz, *Business and society: Ethics, sustainability, and stakeholder management*: Cengage Learning, 2014.

[15] L. Richards, *Handling qualitative data: A practical guide*: Sage Publications, 2009.

[16] G. R. Gibbs, *Analysing qualitative data*: Sage, 2008.

SESSION

BIG DATA ANALYTICS + DATA WAREHOUSE DESIGN + SOCIAL MEDIA + BUSINESS INTELLIGENCE AND KNOWLEDGE DISCOVERY + RECOMMENDER SYSTEMS

Chair(s)

TBA

Supporting crowd-funded agile software development projects using contingent working: Exploring Big Data from participatory design documentation and interviews

J. Bishop

Centre for Research into Online Communities and E-Learning Systems
Ty Morgannwg, PO Box 674, Swansea, SA1 9NN

Abstract - *Designing an effective organisational architecture for an undertaking can be considered essential to its success. The way an organisation is designed – or otherwise appears to its workers – will affect the extent to which those workers associated with it can be effective at their jobs. This chapter undertakes a case study using Big Data from a project called "QPress" that was run by an organisation that is based around contingent working and inter-professionalism. Important things drawn from the data collected from the study include the importance of the Cloud to distance working, such as teleworking; the identity of the organisation and how workers relate to it; as well as what factors assist or inhibit worker motivation. The study concludes that the organisational structure of the organisation investigated – where different firms perform different tasks could be seen as best practice in supporting inter-professional environments.*

Keywords: Big Data, organisational architecture, contingent working, project management, social audits

1 Introduction

The design of organisations has often followed a hierarchical process, where there is a single leader at the top from whom all others' authority is derived. Such a mechanism is lacking in many regards, particularly in inter-professional contexts. The idea that creativity can be fostered in an environment where each person is under the command of another does not follow. The concepts of composite and contingent working were of much interest in the second-half of the 20th century and the early years of the 21st century [1-4], but were of less interest 2005-2010, where the UK Government launched an attack on the construction industry that was almost entirely contingent, even if still hierarchical. In 2007 the UK trade unions showed its grip on the Labour Government by getting them to bring construction workers onto the employee payroll, through using the power of the tax authorities to find instances where proper sub-contractor contracts did not exist, dressing this abuse of power up as dealing with "tax avoidance." By destroying the contingent approach to working in the construction industry the government caused many construction firms to become insolvent during the 2008 global recession, as they were paying the wages of people for whom they had no work. The UK Government's plaster to fix the mess they caused in the construction industry was a scheme called 'ProAct.' This meant that those construction firms, who kept on workers they otherwise couldn't afford or offer work to, would in exchange for keeping them on for fewer hours, get government funding to retrain during the recession. This chapter aims to show how if the successes of the construction industry when it was contingent are applied to the software development industry, then concepts like agile programming, inter-professionalism and crowd-funding can become successful realities.

1.1 Organisational Architecture and Organisational Learning

In the same way an architect can describe the entire building using a blueprint or drawing, organisational architecture is a document that outlines the holistic works (largely unseen) of the organization [5]. According to most authors writing about organisational architecture, the concept is a metaphor, as traditional architecture determines the form of the institutional space where life will be led. Organisational architecture is often considered the bridge between the strategy of an organisation and its workflow processes, helping to convey a consensus about a unique picture of the organisation [6].

Organisational architectures are often thought of in terms of hierarchies and management paths, and the idea of businesses being learning communities with worker-led education is often an alien construct [7].

In France, the expression "theory of organisational 'architecture'" has already begun to progressively replace the expression of "theory of corporate governance" [8]. In developing new organisational architectures, it is considered possible to adopt stakeholder-specific values, which has particular relevance when organisational architecture is complex and the needs of significant other stakeholder groups need to be taken into account [9]. An important aspect of organizational architecture is where the organizational structure has been designed and developed to facilitate the process of knowledge creation [10]. This is probably no more important than in organisations dependent on contingent working, where it is often the case that the workers that

undertake tasks for it depend on generating their own knowledge and practices more so than depending on those working for the organisation engaging them [10].

1.2 Contingent Working and Inter-professionalism

In crowd-funded projects that have unstable revenues, contingent working can be used as a means of hiring employees for tasks only at the points in the project where there is funding available. There is no clear and unambiguous definition of contingent working. Some authors have argued it is a period of self-employment, full or part-time employment of less than one year for a change of job status [11], but this does not consider freelancers and those working under other zero-hour contracts. Others define it as a situation where an employee is taken only when needed [12], which is generally how the author conceptualizes practice. Some argue that contingent work is synonymous with temporary working or short-term employment [4, 13], but this ignores the fact that contingent workers are available for a particular company on a permanent basis.

The use of contingent work is modest, but increasingly decisive [3]. In the UK, two groups of highly skilled, unionized workers where contingent work is common include those working in UK higher education institutions (HEIs) and artists in the entertainment industry [13]. However, contingent work is not universal and logistics companies surround themselves with concentric circles of workers who are more or less central to their operation [14]. Nevertheless, contingent work arrangements are becoming commonplace in the United States at the same time as the number of temporary workers, part-time workers and contract workers continue to expand [15]. Although some workers choose a contingent work status for family or other personal reasons, a large and growing percentage of these contingent workers in America choose these positions as their only option [16]. Even Japanese companies, which are often the most paternalistic, are now using contingent working to reduce fixed costs [12].

2 A Case Study of Crocels and the 'Digital Classroom of Tomorrow' project.

Crocels is the name given to a number of companies and project-based structures that have evolved as a legal corpus over a number of years. It existed as a concept since 2005, becoming 'The Centre for Research into Online Communities and E-Learning Systems.' The creation of a co-operative social enterprise called 'Glamorgan Blended Learning Ltd,' gave the organisation a legal personality, with this company being renamed 'Crocels CMG CYF' (GBL) following the creation of the 'Crocels Community Media Group' (CMG) in 2011. It was then restored to its original intended name, Centre for research into Online Communities and E-Learning Systems (Wales) Limited, following plans for CMG to become a legal entity in its own right bring approved. The creation of CMG as part of GBL in 2011 severed the organisation from the employee and multiple director model required by Co-operatives UK for it to remain a co-operative, and instead became wholly based around those working for it being separate from the organisation, with only one director as permitted by the Companies Act 2006. In 2015 CMG will become the corporate director of all the companies that are a member of, and will allow for the restoring the co-operative status and also lead to a working agreement between all participating organisations and appointed the trustees on its board. Many of the functions of this contract had to still be carried out by GBL, where there would be a risk the activities would not be perceived as charitable in nature. Other functions were transferred to the other companies.

Crocels's organisational architecture is thus a complex one, but it would seem from this study, one that works. The organisation has no employees on the payroll, but the workers could be considered employees using criteria set by the Office of National Statistics. Each worker is a contingent worker, called upon only when their skill is needed. Unlike zero-hour contracts, the participation of workers is totally voluntary, meaning Crocels is more of a co-operative than a workers co-operative which uses employee contracts. These workers also work remotely from a place of their choice, requiring only an Internet-connected device. One of the projects the staff have been working on at Crocels is an agile development project where contingent workers with different skills have been involved at different points in the project. Called the "Digital Classroom of Tomorrow," a number of software products have been and are being independently developed, namely PARLE [], PAIGE, Paix, Vois, MEDIAT, and QPress. Table 1 sets out the original purposes of the DCOT Project, which has been in existence since 2002, and forms part of the Classroom 2.0 initiative [].

Table 1 Objectives of the Digital Classroom of Tomorrow Project

Factor	Objective
Persuasive	Encourage the development of bilingual educational and training materials relevant to Wales.
Adaptive	Develop flexible modes of provision tailored to the needs of the individual learner
Sociable	Strengthen distance learning and use ICT to move away from rigid timetables and classroom based teaching
Sustainable	Encourage the development of essential ICT skills throughout local communities

2.1 Methodology

This study presents the views of those who work for Crocels on the QPress application to attempt to devise general principles of which organisational architectures can best contribute to agile and crowd-funded software projects. The reflexive process used in this study made the structure of Crocels as it

has become much easier to conceptualise, and its unique qualities easier to extract to form general principles and specific directions.

The methodology can be best seen as an extension of ethnomethodology in that it is not only the 'methods' of the group that are identified and discussed, but other elements also. In this case this includes their rules, amities, memes, and strategies as in [17], as well as an additional one; namely enmities, based on 'detachments' [18]. The study was based on the ecological principle, which in contrast to the society fallacy argues that by eliminating the commonalities of those within a group, generalisations can be drawn in terms of the individual and their place within a sample. The society fallacy wrongly assumes that by understanding a small group of people, their commonalities can be used to assume people from a much wider stratum must share those characteristics.

In the case of this study, a number of approaches were used to establish how the ecological principle is reflected within those involved with the Digital Classroom of Tomorrow project. This involved a mixed methods approach and a thematic analysis. In the case of the former, numerical information was used to establish the similarities and differences between the participants by requiring them to provide a 'yes' or 'no' answer to how the organisation is affecting them. The qualitative data generated from them was then used to explain why some members of the group were 'deviant' from the others. This 'deviance,' or more appropriately 'difference,' allows for the ecological principle to be followed in terms of identifying that person's individual identity.

2.2 Participants

The study participants were in three groups – users selected to derive scenarios for the design of QPress, contingent workers to assist with various aspects of QPress's development, and users who tested QPress to see if it met their requirements. In terms of numbers of participants, three were selected. According to standards for IT research, a total of 3 participants is perfect for follow up studies, those involving iterative design and formative evaluations [19]. Choosing 3 participants was also important for ensuring the ecological principle was followed – any even number could mean it would be difficult to eliminate commonalities if it is possible for commonalities and differences to have the same quantity.

The participants selected have had their name changed for privacy reasons. Tom's role was to put together the content for a test of QPress that would seek the involvement of people with autism and those without. Tom was one of the original workers at Crocels, prior to its reform based on contingent working, which he was not too pleased with, as it resulted in his contract not being renewed. Argie was responsible for ensuring the correct documentation was available to team members and that they were up to date with industry practice, such as providing an article on copyright laws from CILIPupdate. Argie was one of the newest workers, who joined Crocels after the contingent working reforms had come in. Jimmy was a programmer, responsible for translating the directions of the project manager into the code to enable QPress to be implemented. Jimmy was fully independent of Crocels, being part of an unconnected firm he was employed by. However, he was a contingent worker like the others, because he was engaged as and when his skill was needed.

2.3 A text-numeric analysis of the data

This section presents the outcome of the mixed methods study where Big Data is generated not only through quantitative data, but through qualitative data also. The data was collected as part of a "social audit," which is where the organisation – Crocels – sees whether it is meeting its objects as a social enterprise. This therefore makes it a form of Big Data because it is collected for one use and then used for another.

2.3.1 Methods

A Method in this context is a particular action [20, 21], activity [22, 23] or similar behaviour that can be seen to be commonplace in a group of people, such as an organisation. The investigation found that all those considered said they were getting what they wanted with working with Crocels, were content with the reasons why Crocels had the practices it does, and that working for Crocels will help them achieve their life goals. Table 2 shows the responses of the three people inspected as part of the survey.

Table 2 Asking workers for their opinion the way Crocels does things, such as distance working, using the Cloud, etc.

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	1	1	1
What is happening in your life?	1	1	0
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	1	1	1
What are you doing at Crocels now that can help you achieve other things in the future?	0	1	1

Table 2 shows that on the whole all staff are satisfied with the various Methods used within Crocels to achieve its aims. Slight differences exist in the case of Jimmy, who found that the methods used by Crocels had little impact on what is happening in life. Equally in the case of Tom, how the Methods used by Crocels can help him improve his effectiveness in Crocels was also not a satisfactory factor. Looking at the workers' explanations for this threw up some unexpected results. Argie's reason for indicating approval was: "*Working with the company helps give me a focus in life, such as meeting deadlines or knowing that sometimes you can't just walk away from any bad situation or task and hope it just gets done, with focus it can be resolved or completed in a timely manner.*" Tom's reason for

indicating approval was "Since my reduction in my hours worked for Crocels impinges very little on my life beyond my research interests into EI & ERD." Tom is very experienced in terms of work history, and whilst his job involved deciphering meanings idioms, which he had a background in, because there was a lot of data handling, it did not play to his strengths and he found it difficult doing the work.

Jimmy's reason was *"It's good experience to work with Crocels!"* This would suggest that the fact Crocels has little impact on the lives of Tom and Argie - because they determine their working conditions - then this meets their approval. Similarly, Jimmy answered differently in the interview to his response, indicating he was dissatisfied with Crocels's methods, but also said he was doing the sort of things he would expect to be doing. Jimmy is on the periphery of Crocels, being a contractor independently constructed. On this basis the involvement with Crocels is just one of many projects Jimmy handles, meaning his development is based on personal growth through the satisfaction working on projects for Crocels. Indeed, it is known that distance working offers many of the benefits of working part-time, including a reduced role of a company in a person's life, as they are able to spend more time with family [24]. The contingent working model of Crocels - based on sub-contracting, involves workers doing their work as a service rather than as a job, and has long been seen as an effective way to ensure the financial flexibility of a firm [25].

2.3.2 Rules

In this context a rule refers to the values of a particular culture that restrict or direct their activities [17]. Rules at Crocels include submitting in accounting periods running from the Tuesday closest to the 15th of the month through to the Monday immediately before that Tuesday. Table 3 shows the responses of the participants when considering the effect the Rules associated with Crocels have on them.

Table 3 Asking workers their thoughts on the way they must do things at Crocels, like the monthly accounting period

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	0	1	1
What is happening in your life?	0	1	0
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	1	1	0
What are you doing at Crocels now that can help you achieve other things in the future?	0	1	0

As can be seen from Table 3, all the workers were getting what in terms of why Crocels had the rules it did, the effect to which it helps them get what they want from Crocels, as how Crocels's rules allow them to get what they want from Crocels. Differences came in terms of the effect Crocels's rules had on what is happening in the workers' lives and the extend to which Crocels can help them achieve things in the future.

In the case of Tom, he felt there was little about the Rules he has to conform to at Crocels can help him in his future goals. He said: *"[M]y current main business work has no relevance to Crocels and my academic career is really nonexistent at this time in my life."* This might imply that Tom feels like an 'invisible employee,' where he is disengaged from the mission, feeling like he is in a thankless job [26]. On the other hand, Argie, was satisfied with the rules at Crocels. "Working with Crocels has made me aware that in good situations or bad situations, you just have to keep focused and carry on, because things can't always go as you hope," he said. "This does not detract from working for Crocels, which has been pleasant."

2.3.3 Amities

Amities in this context are the friendships and other positive relationships that exist within an organisation or culture [17]. In the case of Crocels, this includes opportunities for building positive relationships and maintaining a work-life balance to have out-of-work friendships.

Table 4 Asking workers their views on opportunities for building positive relationships in Crocels, and maintain a work-life balance to have out-of-work friendships

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	1	1	1
What is happening in your life?	1	1	0
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	1	1	1
What are you doing at Crocels now that can help you achieve other things in the future?	1	0	1

As can be seen from Table 4, almost across the board those inspected said they had positive benefits in terms of the social relationships they had with others within Crocels. The two exceptions were Jimmy, in the case of what is happening in their life, and Argie in terms of how they might achieve future ambitions.

2.3.4 Enmities

Enmities in this context refer to those persons who have an inhibiting effect on one's ability to perform effectively in work, living or study, for instance. At Crocels this includes its teleworking policy, which means workers do not have to work with people they might not like. Table 5 shows the respective views of the Crocels workers selected for participation.

Table 5 A workers' views on Crocels's teleworking policy means they don't have to work with people they might not like

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	1	1	0
What is happening in your life?	0	1	1
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	1	0	0
What are you doing at Crocels now that can help you achieve other things in the future?	0	1	0

Most dissatisfaction in this category related to why things were happening at Crocels. Argie was particularly critical of those enmities affecting his job. *"I answered […] unhelpful, because I am usually kept waiting several days past the due date for time sheets to sort the hours and fill in the overall time sheet,"* he said. *"Otherwise, everything else is helpful."* Jimmy on the other hand said: *"Most of the good stuff happened which we found helpful at several stages,"* which suggests that it can be problematic making progress on the software - due to the workload of the software designer.

2.3.5 Memes

In the case of Crocels, memes include prescribed ways to design information systems, and in working to agreements. Table 6 shows the responses to the category from the participants.

Table 6 Beliefs you may be expected to adopt or conform to at Crocels, such as in the design of information systems, or in working to agreements.

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	0	1	1
What is happening in your life?	0	1	0
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	1	1	0
What are you doing at Crocels now that can help you achieve other things in the future?	0	1	0

As can be seen from Table 6, in terms of the culture of Crocels that shape its belief system takes, Argie was satisfied across the board. Tom and Jimmy both indicated that the memes that exist within Crocels do not impact on what is happening in their life, nor influence their achievement of future goals. An important feature of this category is that Argie and Jimmy were getting out of Crocels what they wanted, but Tom was not. *"The constraint placed [by Crocels] is one of threat,"* he said. *"comply or take the consequences,"* he added. *"[T]hose of displeasure at least; abuse at worst."* This would suggest that the premise of the sub-contracting model - delivery else you will not get paid - is not one that Tom finds satisfying. Jimmy on the other hand was completely satisfied with what he was getting from Crocels. *"[T]his is the real experience that we experienced while working with."*

2.3.6 Strategies

Strategies. In the case of Crocels, this includes its mission – based around e-learning and promotion of social change and world peace – where the attitudes of workers are expected to reflect the will to achieve such outcomes. Table 7 sets out the differences of opinion of workers in the ways in which Crocels's strategies affect them.

Table 7 Whether the Crocels' mission – based around e-learning and promotion of social change and world peace – reflects or outlook or whether you have enough opportunity to take part in or improve the mission.

Factor	Tom	Argie	Jimmy
Are you getting what you want from Crocels?	0	1	1
What is happening in your life?	0	1	0
What could you do better at Crocels or because of Crocels?	1	1	1
Why things are happening at Crocels?	0	1	1
What are you doing at Crocels now that can help you achieve other things in the future?	0	1	1

As can be seen from Table 7, differences were found where Argie found the goal-setting process of Crocels helpful in terms of the effect on his life, whereas the others did not. In fact the only area where Tom was satisfied was in terms of how the strategies of Crocels can help increase his productivity in his job, which as discussed earlier were likely to be because Crocels had minimal impact on his life. Argie found that the organisation's mission and strategies was helpful. "The work (at Crocels) is relaxed, but always done in time, and aside from one person being late with time sheets, everything is satisfactory," he said. "It has been argued that the goal-setting process is ineffective, especially because changes in organisational goals, if not communicated, will not be seen as a series of minor changes when they are eventually picked up by workers used to working a specific way" [27].

2.4 A thematic analysis of the data

This study aimed to draw out themes from the two sets of interviews, which can be used to audit the effectiveness of Crocels's organisational architecture. Findings includes that

teleworking and Cloud-based computing are an important factor, as is Crocels's corporate identity. Finally the motivation of staff was a key issue, with many findings been drawn out of that.

2.4.1 Teleworking and Cloud-based Computing

It has been argued that teleworking is not suitable where there is a high visibility element to work, interfaces with clients are essential, where there is a high creative element or where there are risks or hazards that need to be managed [28]. However, such claims have been made prior to the Internet era, and the contingent working philosophy of Crocels is ideally suited to teleworking. This was evident in the data, and the addition of the Cloud into the way Crocels now does things – going beyond email – is appreciated by the workforce. Teleworking is best described as a flexible approach to working where activities are carried out remotely, involving telecommunications in one form or another [29].

2.4.2 Corporate identities

It has been argued that learning is about developing an identity and becoming a practitioner rather than the objective and prescribed approach used in teaching [7, 30]. This has been important to the development of Crocels as a complex organisation on paper, but intended to be a simple one to those associated with it, through the creation of a single branding, albeit of a number of years of reflective practice. As 'Argie,' put it: "*Working with Crocels has made me aware that in good situations or bad situations, you just have to keep focused and carry on, because things can't always go as you hope. This does not detract from working for Crocels, which has been pleasant."* This shows that the idea of an organisation as a 'nexus of contracts,' which reflects a corporate autonomy, is not suited to learning organisations [31]. Indeed, the aim of Crocels has been for its workers to be self-directing managers of their own work, which the 'contingent working' model has helped make possible. This means they will be responsible for a lot of 'self-organised learning,' through which the personal construction of meaning in a learning organisation is the basis on which actions within that body are based, so as to create a sense of 'personal knowing' [32].

2.4.3 Motivation

Every organisation needs a leader, whether or not this person is part of the day-to-day operations of the organisation, such as Richard Branson in Virgin who takes the role of a mentor [33], Bill Gates, formerly of Microsoft, who was more strategic [34], and Steve Jobs at Apple, who provided certainty in terms of expectations [35]. Crocels is no different, and the CEO does, and should, take on many of these roles as and when required.

Poor CEOs provided few opportunities for learning, and through being insincere in their wish to 'empower' workers can often demotivate then when it is not put into practice [36]. But this is not the case with Crocels, and as 'Argie' said, the contingent working and self-management approach was suited to him. "Working with the company helps give me a focus in life, such as meeting deadlines or knowing that sometimes you can't just walk away from any bad situation or task and hope it just gets done, with focus it can be resolved or completed in a timely manner," he said.

It is clear, however, that for contingent working to be empowering, it has to be assigned to the right person, with the right skills, at the right time. Tom for instance described his work as having "collected a range of text & some graphics source materials," for the AHI project, but found these "demanding, boring, mundane, frustrating/annoying," and even with his specialist knowledge of idioms, he found "language skills demanding."

The dislike of manual or routine work by Tom had implications for other workers in other areas of the company, as Argie clearly articulated. "Working for Crocels gives me everything I could want from working for the company," he said. "The work is relaxed, but always done in time, and aside from one person being late with time sheets, everything is satisfactory." However, Tom was one of the original members of the company, before the contingent model was agreed to be part of the organisational architecture in 2011.

3 Towards a 'network of practice' approach to organisational architecture and learning

In terms of promoting learning and cooperation, one might think of the Crocels approach as being both pedagogically and otherwise philosophically 'constructivist,' even though it might not appear that way to those working within it. Constructivism in a pedagogical sense states that learning is best achieved through those who can do something rather than working with those who cannot [37, 38]. In terms of a philosophy of science, it states that a person's view of the world – often seen through epistemology and ontology – is constructed by them through their own senses and is different to others both in terms of the words and concepts they use and the meanings given to them [39]. Social theories of learning often try to challenge the solidary or one-on-one competitive drive that exists in many people looking for a balanced life, with collaboration as opposed to cooperation being the key factor [40]. At Crocels things are different, and learning can be thought more effectively as 'networked' and the company thus a 'network of practice.'

It has been argued that in an interconnected world as exists now, social bonds need not be as strong as that required for situated or social learning, as people are now befriended for the uses and gratifications they provide no different to any other mediated entity [41]. Called 'networked individualism,' this on-going phenomena, since David Kemp |first described networks beyond frontiers [42], means the strong bonds needed for traditional relationships can be given way to more manageable relationships with more people than possible otherwise [41].

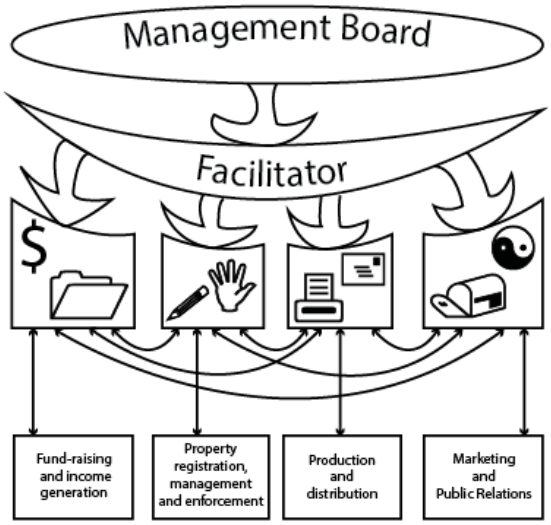


Figure 1 A generalised framework for an agile software development company

A concept of 'networks of practice' to refer to an organisation like Crocels, and 'networked learning' to refer to the process by which members of an organisation learn within the social context of what they are doing and who they are with at any given time, is more suited in a time when networked individualism is the norm. It has long been known that individuals construct the world through the assigning of symbols or signs to link external representations with internal ones [37, 43]. On this basis, even though it is possible for the author to conceptualise Crocels's structure through Figure 1, it will likely have been constructed differently by each person participating with the organisation, and thus lessons learned from it will need to be re-constructed to apply to other organisations. Figure 1 sets out an example of how 'networked learning' is represented as an organisational architecture in terms of Crocels, including the pathways between the companies showing how 'networked individualism' can occur. It can be seen that the four existing companies are specialised around fund raising and income generation; property registration, management and enforcement; production and distribution; and marketing and public relations. In addition on top of these is a management board and a facilitator. The facilitator is the director of all the organisations and the CEO of the group (existing through the management board), through which all decisions are approved, or indeed rejected, based on their individual responsibilities and obligations to each company. At the heart of this design are the important factors of competence, risk management and organisational efficiency. Table 8 sets out the competencies, risk management and organisational efficiencies require for this model to be realised. It has been argued that statements of competence are an effective tool for the development of individual staff, which enables learning organisations to provide a structure that can assist in the planning, delivery and evaluation of all personnel development activities [44]. Indeed, as discussed earlier this was important to Argie and having deadlines and similar consistent processes increases his satisfaction.

Table 8 Specific purposes for companies within the Crocels Community Media Group

Company Purpose	**Staff competences**	**Risk Management**	**Organisational Efficiency**
Fund-raising and income generation	Marketing, initiative, stakeholder management	Public liability	Even if funding is channelled through the Group, separating financial and public-facing risk from it allows for a greater focus on doing.
Property registration, management and enforcement	Legal knowledge	Contents liability. Having all the property held in a company with no customers beyond the other companies in the Group ensure the assets for all are protected.	By not being responsible for the creation or development of IP then it is subject to less risk and lower insurance costs
Production and distribution	Logistics, customer service, project management	Products liability. Having one company and a set of subsidiaries in each territory responsible for development	By focussing on developing on the research licensed from the other firm, it can take more risk in developing products without the costs other firms take on.
Marketing and public relations	Marketing, Internet searches, stakeholder management, persuasion and negotiating.	Legal expenses cover, such as for risk of mis-selling and defamation	By focusing on promoting the Group's products and services, the risks are felt by other companies, allowing focus on realising IP.

In terms of risk management, for an organisation as complex as Crocels, if it existed as a single legal person acquiring insurance might be impossible, as the risk to be ensured is significant. Thus breaking the organisation into various autonomous

specialist units can best facilitate Crocels's activities and its capacity as a learning organisation.

Efficiency is known to be an important part of reducing the running costs of an organisation and increasing customer satisfaction. Crocels has done this in a number of ways, including developing a single identity for its client facing websites, whilst also having corporate websites for each company and the group as a whole. The intention is that the complexities of various specialist organisations is not obvious to customers, and they access Crocels's goods and services through the website optimised for their region rather than each separate organisation's website.

4 Implications and Future Research Directions

This study has looked at the importance of establishing an organisational architecture that is based on learning, should one want to build an inter-professional environment where the personal development of workers is of paramount importance. The study found that an organisation's intended and actual culture has an impact on the extent to which workers feel part of it and their future plans enabled. Through using the 'M-REAMS' approach to ethnomethodology, the study was able to pinpoint the reasons why workers expressed satisfaction or dissatisfaction with the way the organisation investigated – Crocels – was run. Future research should look at whether the findings of this study – based on three contingent workers, is applicable more generally with other organisations that take part in a flexible working approach to organisational architectures.

5 Discussion

Designing an effective organisational architecture for an undertaking can be considered essential to its success. The way an organisation is designed – or otherwise appears to its workers – will affect the extent to which those workers associated with it can be effective at their jobs. This chapter undertook a case study into an organisation – called Crocels - an organisation based around contingent working and inter-professionalism. The worker environment at Crocels supports a composite approach to working, where workers are their own managers, and their work reflects what needs doing as opposed to simply doing the same job endlessly and doing nothing when a job outside of one's job description needs doing.

Important conclusions drawn from the study include the importance of the Cloud to distance working, such as teleworking; the identity of the organisation and how workers relate to it; as well as what factors assist or inhibit worker motivation. The extent to which workers feel a part of an environment is essential to that organisation's success, and the study shows that those workers who felt left behind by organisational change are the least satisfied. The study also found that those new members of Crocels were more enthusiastic about the change the older members were exposed to, but which the newer members had little awareness of.

The study concludes that the organisational structure of the organisation investigated – where different firms perform different tasks, could be seen as best practice in supporting inter-professional environments.

6 Annex - Interviews with Workers

Table 9 shows the interview questions given to staff, and their answers, in relation to the organisational architecture of Crocels.

Table 9 Interviews with workers on the organisational architecture

Question	Tom	Argie	Jimmy
Please discuss whether you are getting from Crocels the things you want to.	There is not the range of options I would choose in the answers above. Areas like Methods, Rules & Strategies can seem unclear because they are not always apparent in this way of working. They will seem undefined and 'open', until questioned, and only then considered and then dismissed (e.g. I was not aware 'world peace' featured in Crocels mission). Enmities section is simply confusing. You offer an option in the question to reverse your choice options!? You should perhaps simply asking something like - 'Do you find it easier working with Enmity challenging personalities or without..?' The constraint placed (as with these surveys) is one of threat - i.e.'comply or take the consequences': those of displeasure at	Working for Crocels gives me everything I could want from working for the company. The work is relaxed, but always done in time, and aside from one person being late with time sheets, everything is satisfactory.	Because this is the real experience that we experienced while working with

	least; abuse at worst.		
Please discuss how things happening in your life are affected by your participation in Crocels activities	My answers above applies here too. Since my reduction in my hours worked for Crocels impinges very little on my life beyond my research interests into EI & ERD.	Working with the company helps give me a focus in life, such as meeting deadlines or knowing that sometimes you can't just walk away from any bad situation or task and hope it just gets done, with focus it can be resolved or completed in a timely manner.	It's good experience to work with Crocels!
Please discuss how Crocels is helping make your life better.	Much in my first answers above applies here too. I should also say that my work & study motivations are personal & private.	Working through the cloud saves time, because files are shared instantly with those who need them. This makes it easier to carry out tasks, and the use of skype keeps the need to contact others working for the company.	It does make it better the way you incorporate your self within.
In terms of the events in your life that occur through your participation in Crocels activities, how helpful or unhelpful are these to you?	Since my reduction in my hours worked for Crocels has very little effect on my life beyond my research interests into EI & ERD. I should also say that my work & study motivations are personal & private.	I answered the third statement with unhelpful, because I am usually kept waiting several days past the due date for time sheets to sort the hours and fill in the overall time sheet. Otherwise, everything else is helpful.	Most of the good stuff happened which we found helpful at several stages.
Please discuss how Crocels helps you achieve your life and career ambitions and what about Crocels might get in the way of those.	Much in my first answers above applies here too. Also my current main business work has no relevance to Crocels and my academic career is really nonexistent at this time in my life.	Working with Crocels has made me aware that in good situations or bad situations, you just have to keep focused and carry on, because things can't always go as you hope. This does not detract from working for Crocels, which has been pleasant.	Glad to share our thoughts!

Table 10 shows the responses to interview questionnaires by the workers selected to take part in the study.

Table 10 Responses to questions on contingent working

Question	**Tom**	**Argie**	**Jimmy**
1. If you took part in discussing the suitability of given sites in terms of style, ease of use and the end user in order to provoke reactions or responses in others please comment on your experience here:	I have taken part in discussions about issues relating to idioms, aphorisms and/or cliches in commonly used imagery phrases, as well as graphics for storyboards. My general experiences when discussing the VOIS system were demanding, interesting & frustrating. However, the design of VOIS clearly lacks features and considerations needed to achieve what it set out to do.		
4. If you took part in collecting a range of source materials (e.g. text, graphics) that can be used and modified in various contexts to provoke reactions or responses in others please comment on your experience here:	I have collected a range of text & some graphics source materials. These took the form of spread sheet data of idioms, aphorisms and/or cliches and commonly used imagery & phrases, as well as hand & computer graphics for use as story-boards. Experiences were: - Text: technical (demanding, boring, mundane, frustrating/annoying) & language skills demanding. Graphics: demanding of drawing & composition skills.	Very easy to find such material with the programs/ sources used, and provided a lot of material to be used.	
5. If you took part in selecting the	.	Checking through magazine	

most appropriate materials, considering factors like file size and saving files in appropriate format in order to provoke reactions or responses in others please comment on your experience here:		articles, it took time to find articles of interest which could then be used later if need be. This has provided less, but no less relevant material	
6. If you took part in storing files in a preparation folder (e.g. DropBox or web servers) that are intended to provoke reactions in others please comment on your experience here:	I did this on my lap-top(HP7000) computer here as well as, as a folder in my secure 15G Google Drive web storage space.	Using Dropbox to store files has been a fantastic time saver for the tasks given to me. It has allowed me to upload and store all files required in a matter of seconds, providing an indefinite backup of everything .	
10. If you took part in creating text and images to provoke reactions or responses in others please comment on your experience here:	I have created a range of source materials in the forms of idioms, aphorisms and/or cliches in commonly used imagery & phrases, as well as graphics for use as storyboards. Experiences were: - Text: technical (boring, mundane, frustrating/annoying) & language skills demanding. Graphics: demanding of drawing & composition skills.		
17. If you took part in collecting or giving feedback, such as in terms of suitability for purpose, ease of use and style, please comment on your experience here:	My experiences when discussing the VOIS system were demanding, interesting & frustrating, as I've said. VOIS is conceived as a fully automated ER system based on a semantic intranet where ERD users can enter a 'mug-shot' inquiries and get answers. From their experience they can update and improve the system's intranet database of answers. When the priorities of VOIS changed to considering wider, more general security & verity assessing applications, developments of the original purpose of the system seemed sidelined.		

7 Acknowledgements

The author would like to acknowledge all those who commented on earlier versions of this paper. Results from the "QPress" project were presented at WordCamp Birmingham 2015 at The Studio in Birmingham on 8 February 2015, where novel findings on crowdfunding projects, such as the links between countries and user motivations, were first presented [45].

8 References

[1] P. G. Herbst. "Autonomous Group Functioning: An exploration of behaviour theory and management". Associate Book Publishers Limited, 1962.

[2] Joel Krieger. "Undermining Capitalism: State Ownership and the Dilectic of Control in the British Coal Industry". Pluto Press, 1984.

[3] Jacqueline A-M Coyle-Shapiro & Ian Kessler. "Contingent and Non-Contingent Working in Local Government: Contrasting Psychological Contracts"; *Public administration,* 80., 1, 77-101, 2002.

[4] Surhan Cam, John Purcell AND Stephanie Tailby. "Contingent employment in the UK"; *Contingent Employment in Europe and the United States* (Edward Elgar Publishing) Ola Bergström & Donald W. Storri (Eds.), 52-782003.

[5] GJ Lee, R. Venter & B. Bates. "Enterprise-based HIV/AIDS strategies: Integration through organisational architecture."; *South African Journal of Business Management,* 35., 3, 2004.

[6] Jorge S. Coelho. "How to align information systems requirements? MLearn approach". 7th International Conference on the Quality of Information and Communications Technology–Industrial Track, 2010. .

[7] Josh Plaskoff. "Intersubjectivity and community building: learning to learn organizationally"; *The Blackwell handbook of organizational learning and knowledge management,* 161-184, 2003.

[8] Dominique Bessire. "Transparency: a two-way mirror?"; *International Journal of Social Economics,* 32., 5, 424-438, 2005.

[9] Richard Hudson. "Brand strategy for acute NHS trusts"; *Journal of Communication in Healthcare,* 2., 1, 20-33, 2009.

[10] Susan Durbin. "Knowledge Creation, Call Centres and Gender–A Critical Perspective"; *Brain Drain Or Brain Gain?: Changes of Work in Knowledge-based Societies,* 14., 241, 2011.

[11] Stephen Gorard. "Patterns of work-based learning"; *Journal of vocational education and training,* 55., 1, 47-64, 2003.

[12] Christopher Davis. "Maintaining a Balance?"; *I in the Sky: Visions of the Information Future,* 18, 2000.

[13] R. Cunha, M. Cunha, ANTÓNIO Morgado & Chris Brewster. "Market forces, strategic management, HRM practices and organizational performance, A model based in a European sample"; *Management Research,* 1., 1, 79-91, 2003.

[14] Roger Sealey. "Logistics workers and global logistics: The heavy lifters of globalisation"; *Work Organisation, Labour and Globalisation,* 4., 2, 25-38, 2010.

[15] Jamie Peck & Nik Theodore. "Contingent Chicago: Restructuring the spaces of temporary labor"; *International Journal of Urban and Regional Research,* 25., 3, 471-496, 2001.

[16] Robert E. Parker. "The Global Economy and Changes in the Nature of Contingent Work"; *Labor and Capital in the Age of Globalization: The Labor Process and the Changing Nature of Work in the Global Economy,* 107-123, 2002.

[17] Jonathan Bishop. "The Equatrics of Intergenerational Knowledge Transformation in Techno-cultures: Towards a Model for Enhancing Information Management in Virtual Worlds". 2011. .

[18] Jonathan Bishop. "The role of the prefrontal cortex in social orientation construction: A pilot study". Poster presented to the BPS Welsh Conference on Wellbeing, Wrexham, GB. 2011. .

[19] Jakob Nielsen. "Usability Engineering". Academic Press, 1993.

[20] L. A. Suchman. "Plans and Situated Actions: The Problem of Human-Machine Communication". Cambridge University Press, 1987.

[21] L. A. Suchman. "Human-machine reconfigurations: Plans and situated actions". Cambridge University Press, 2007.

[22] Y. Engeström. "Activity theory and individual and social transformation"; *Perspectives on activity theory,* 19-38, 1999.

[23] Y. Engeström & R. Miettinen. "Perspectives on activity theory". Cambridge Univ Pr, 1999.

[24] Susan McRae. "Flexible Working Time and Family Life: A Review of Changes". Policy Studies Institute, 1989.

[25] Alastair Evans & Les Walker. "Sub-contracting"; *Flexible Patterns of Work* (Institute of Personnel Management) Chris Curson (Ed.), 143-1651986.

[26] Adrian Gostick & Chester Elton. "The Invisible Employee: Using Carrots to See the Hidden Potential in Everyone". John Wiley & Sons, 2010.

[27] Jim E. H. Bright & Robert G. L. Pryor. "Goal Setting: A Chaos Theory of Careers Approach"; *Beyond Goals: Effective Strategies for Coaching and Mentoring* (Gower Publishing Ltd) Susan David, David Clutterbuck & David Megginson (Eds.), 185-2092013.

[28] John Stanworth & Celia Stanworth. "Telework: The human resource implications". Institute of Personnel Management, 1991.

[29] Mike Gray, Noel Hudson & Gil Gordon. "Teleworking Explained". John Wiley & Sons Ltd, 1995.

[30] John Seely Brown & Paul Duguid. "Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation"; *Organization science,* 2., 1, 40-57, 1991.

[31] John Child & Suzana Rodrigues. "Social Identity and Organizational Learning"; *Handbook of Organizational Learning and Knowledge Management* (Blackwell) Mark Easterby-Smith & Marjorie A. Lyles (Eds.), 305-3292011.

[32] Sheila Harri-Augstein & Ian M. Webb. "Learning to change". McGraw-Hill International (UK) Limited, 1995.

[33] Richard Branson. "Business Stripped Bare: Adventures of a Global Entrepreneur". Virgin Books, 2008.

[34] Bill Gates. "Business@ the speed of thought"; *Business Strategy Review,* 10., 2, 11-18, 1999.

[35] Jeffrey S. Young & William L. Simon. "iCon Steve Jobs". John Wiley & Sons, 2006.

[36] Chris Argyris. "On organizational learning . Malden, MA". Blackwell Publishing, 1999.

[37] L. S. Vygotsky. "Mind in society". Harvard University Press, 1930.

[38] L. Dixon-Krauss. "Vygotsky in the Classroom: Mediated Literacy Instruction and Assessment.". Addison Wesley Longman, One Jacob Way, Reading, MA 01867, 1996.

[39] J. Derrida. "Writing and difference". Psychology Press, 2001.

[40] Etienne Wenger. "A social theory of learning"; *Contemporary theories of learning* (Addington, GB) Knud Illeris (Ed.), 209-2182009.

[41] Barry Wellman & Lee Rainie. "Networked: The new social operating system"; 2012.

[42] D. A. Kemp. "Society and electoral behaviour in Australia: a study of three decades". University of Queensland Press, 1978.

[43] Gavriel Salomon. "Interaction of media, cognition, and learning". Jossey-Bass Publishers, 1979.

[44] Peter Lassey. "Developing a learning organization". Kogan Page, 1998.

[45] Jonathan Bishop. "Crowdfunding WordPress plugins: The case of QPress". The 8th United Kingdom WordCamp (WordCamp Birmingham 2015), Birmingham, GB. 7-8 February 2015.

Data Warehouse Design Using Row and Column Data Distribution

Behrooz Seyed-Abbassi and Vivekanand Madesi
School of Computing, University of North Florida, Jacksonville, Florida, USA

Abstract - *Design of an efficient data warehouse that can provide a comprehensive platform for big data storage and an optimized query processing has been receiving major attention by various organizations. It is essential for the data modeler to have good knowledge and understanding about the design of an appropriate data warehouse based on the structure of the data and how the information is stored for query retrieval or presentation. In this paper, four different methods are presented that extend the design of the star schema and snowflake schema to better support the data warehouse design. The methods were implemented and their advantages and disadvantages based on the designs and query processing performances were analyzed and compared to each other. A designer could use each method individually or in combinations to provide more suitable and efficient data warehouse structure for implementation and query processing.*

Keywords: Data Warehousing, Star Schema, Snowflake Schema, SQL, Data Mining, Normalization, Optimization

1 Introduction

Over the years, Business Intelligent (BI) for Data Warehousing has received major attention from various organizations. It has become very important in storing and retrieving data for decision making. Since the introduction of the data warehouse, the process of building a warehouse as a centralized data collection and repository for Decision Support Systems (DSS) has evolved through phases of research development and is becoming more widely accepted by diverse business organizations. The creation of a serviceable data warehouse requires substantial effort and experience in the intricate design as well as sophisticated implementation procedures. In a data warehouse, the database expands immensely. This combined with the often ad-hoc and complex queries to the database dictates that an essential task in the design process of the warehouse must be to consider ways to reduce the response time and thus minimize the cost of the query effort [1], [2], [3].

In the design of a data warehouse, one of the fundamental structures utilized is the star schema, which has become the design of choice because of its compact structure. The structure utilizes a central table and one level of surrounding tables that results in a more optimized query performance and ease of use compared to the structures of prior data warehouse designs. One structural dilemma that can occur in the star schema involves the handling of multi-values and duplicate values [4]. The effected tables are commonly normalized to a higher level resulting in another structure referred to as a snowflake schema [5]. Although the snowflake structure can handle the issues of multi-values and duplicate values, the higher normalization of the design also increases the complexity of the design structure by adding more levels of tables. The expansion of the number of tables dramatically raises the number of joins required for queries thus prolonging the query retrieval time. The end result is a decrease in the efficiency of the query response time for the DSS.

This paper will offer four design methods with two creative techniques to overcome the star schemas limitations in supporting multi-values and duplicate values with displays of time performance for each method. The techniques maintain the values in the star structure and thereby avoid the need to proceed to the snowflake schema. The preservation of the star schema in the data warehouse design retains the simpler design structure and allows a variety of queries to be efficiently processed. First, the current utilization of Method 1 and Method 2, which are the star and snowflake, will be reviewed. This is followed by a description of Method 3 using a known number of multi-valued attributes and Method 4 using an unknown number of multi-valued attributes. Both Methods 3 and 4 utilize an extended star schema. After the descriptions, the application of efficient data mining using different queries with simple, complex and aggregated structures is examined and the time performances will be analyzed for enhanced decision-making and design strategies.

2 Data Warehouse Design

For a data warehouse, relevant information is gathered from an existing operational database in the form of records and stored in the data warehouse for future query processing, analysis, and evaluation. The collections of records are stored

in two types of tables, one fact table and many dimensional tables [6], [7]. The operational data, grouped into related records as historical data, are distributed into their related dimensional tables according to their types. The fact table holds the keys from each dimensional table. The configuration of the fact table holding a key from each dimensional table to associate all the related dimensional tables together results in a structure called the star schema. A data warehouse may contain many star schemas with each one supporting different dimensionalities through association of related records in the fact table. An example of a star schema is given in Figure 1 that includes Customer, Location, Product, and Calendar to support a fact table for Sales. As shown in Figure 1, the Method 1 schema supports a simple design that allows the data from the operational database to be Extracted, Transformed and Loaded (ETL) to the star structure.

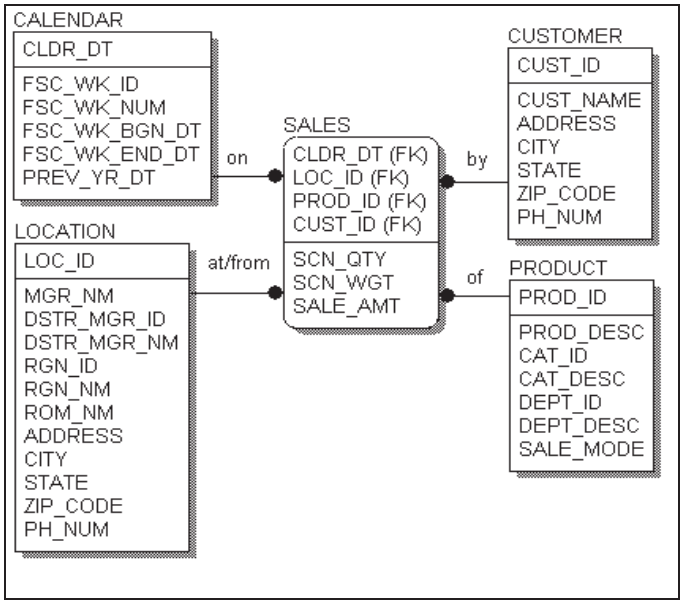


Figure 1. Method 1- Star schema with one fact table and four dimensional tables

When storing information in a star schema, the duplicate data in the dimensional tables may result in unnormalized tables. Generally, duplicate values are handled by the process of higher normalization to decompose the original table into two or more tables with the duplicate values moved to one of the new tables. As an example in Figure 1, the action of normalization may be applied to the tables if the product location in the location table supports multiple district managers and regions for storage locations, if the product has a different category and department, or if the calendar is categorized based on year, month and week grouping.

In this case, the multiple storage locations could be handled by keeping the storage data in a separate dimensional table and associating the key of the location table to the storage table. The new table relation from the fact table can then be viewed as a two level access for retrieval of the storage location information during the join. Similarly in the second and third cases, a more detailed product and calendar can be considered in the design with more normalization to handle various products and dates. The issues of multi-values and duplicate values in the star schema can be alleviated by decomposing the related dimensional tables into more normalized tables, as shown in Figure 2. In this method, the product table may have different locations suggesting various storage areas based on district manager or based on regions instead of having one location for storage or product. If necessary for consideration in the data warehouse system, this type of information (such as multiple locations by district or region for product numbers for each location) can have a higher normalized design as shown in the product table from Figure 2.

The normalization of the star schema from Figure 1 converts the related dimensional tables to a higher normal form table(s) named as Method 2 for design of the structure. The higher normalized star schema structure results in more than one level of tables associated to the fact table. The new structure in Method 2 is referred to as a snowflake schema [8], [9]. This schema is more organized than the star schema in that it reduces the duplicates, but it also results in more dimensional tables from decomposition of original table(s), which in turn requires more complex query joins and lengthens the query response. Although the multi-values and duplicate values may be handled by normalizing a star schema into a snowflake schema, the advantages of the star design for design simplicity and query optimization are sacrificed. The proposed techniques in this paper offer alternative methods of handling multi-values and data normalization while preserving the star schema. Both methods are extensions to the star design and reduce the need to utilize the snowflake structure.

3 Extended Column-Wise Star Schema

In Method 3, a new schema to preserve the structural efficiency of the star schema and to allow multi-value data to be used in the same dimensional tables is provided. The operational tables and their data structures allow the designer to have a good working knowledge of the multi-value fields or duplicate attribute names that are going to be used for modeling the data warehouse. This method will take multiple values for the same attributes from the operational database and move them to a star schema in a column-wise order in the dimensional table(s).

Application of higher normalization to the star schema in Method 1 resulted in the snowflake structure in Method 2. In Method 3, rather than transforming the star schema to a snowflake by utilizing higher normalization, the design is kept in the star schema structure by employing the multiple values, such as storage locations, district managers or regions, with a special arrangement in the location table which is similar to normalization.

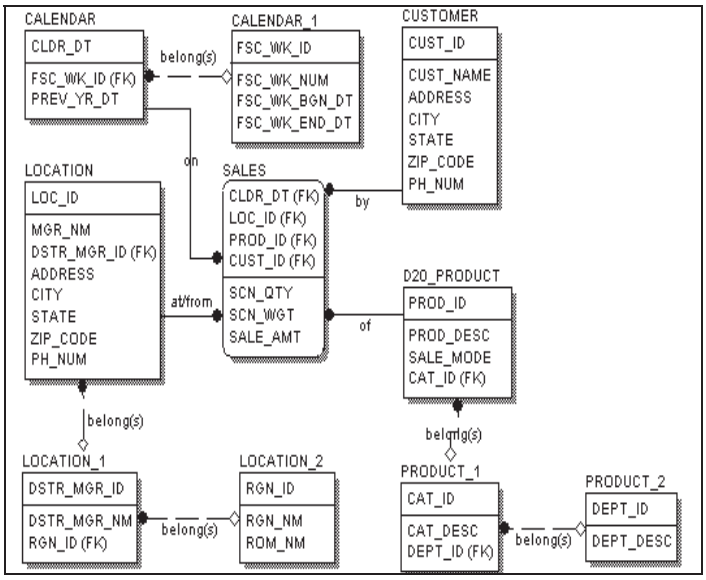


Figure 2. Method 2 - Snowflake schema with one fact table and four dimensional tables

In Method 3, duplicate attributes are distributed column-wise in the same table, instead of normalizing them by moving them to another table or entity. This is also done with other entities, such as product and calendar. Since the designer of the data warehouse knows in advance or ahead of time about the number of the columns for attribute names (such as storage locations, product, and calendar) from the operational database, this knowledge will facilitate the design process. The dimensional tables with multiple attribute values can be created as Method 3 with their new columns to preserve the star schema structure similar to Method 1. Figure 3 shows the star schema with multiple attributes (such as CLDR_DT1, 2, 3, 4, 5, PREV_YR_DT1, 2, 3, 4, 5, ...) for the data warehouse design.

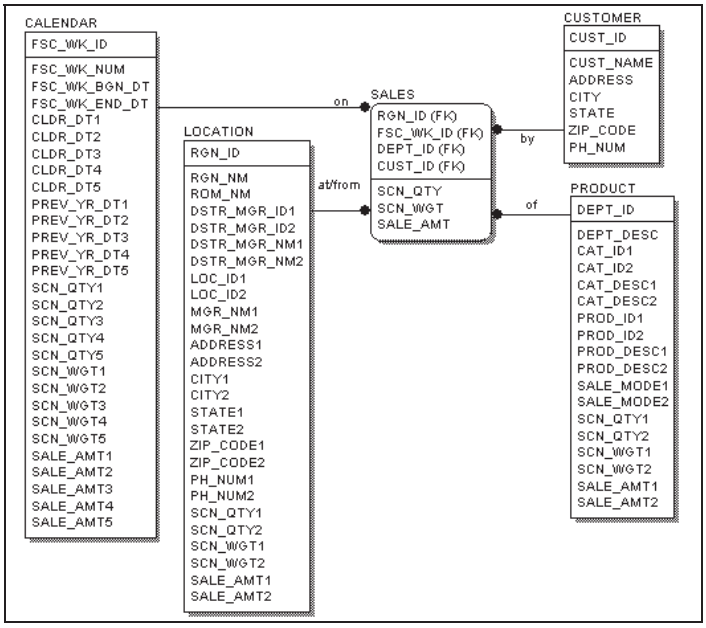


Figure 3. Method 3 - Column-wise distribution

The disadvantage of this method is that some of the records in the tables (such as calendar) may have less number of values for sales amount. This will result in null or no value in that column. On the other hand, the advantage of this method is that the result is a compact design having fewer tables by allowing multiple attribute names and values to be present in one table (such as the product table) as opposed to two or more tables using higher normalization. This method also supports the star schema structure for simplicity and efficiency of the design which results in more optimized query retrieval with less number of joins in the dimensional tables. In general, when the number of multi-value attributes is known from the operational database system, Method 3 is the most suitable design as a technique to maintain the star schema in the data warehouse design.

4 Extended Row-Wise Star Schema

The operation databases are dynamic. Due to organization growth, operational databases expand daily through the addition of new information. This information could include new products or locations to the provided example as a result of expansion. Filtering of these types of multi-value data for the historical data will result in higher normalization and a structure similar to the snowflake schemas in Method 2. When the number of duplicate attributes or possible multiple values for the tuples of the data warehouse from different tuples of the operational databases are not known, Method 4 can be utilized in the data warehouse design to preserve the star schema structure by avoiding the snowflake schema and decomposition of the dimensional tables with duplicates into two or more tables.

In Method 4, the data warehouse can be designed using the historical data from the location table and new location by categories or managers allowing for future expansion of historical data to be handled through distribution of multiple locations in a tuple of the dimensional tables rather than the attributes of the location dimensional table. As shown in the following schema of the location dimensional table, the district manager and region attribute names will be repeated only once. This method will allow for future expansion of a new manager data and district by adding the new historical tuple to the location dimensional table. In Method 4, any multi-valued attribute is added as new record with a Row-Id to the table (existing table) instead of being normalize by moving to anther table. To recognize all the records with the same RGN_ID, a new attribute with the name LOC_ROW_ID is used to represent the location records with duplicate values that are added to the data warehouse table. An example of Method 4 is shown in Figure 4.

The advantage of this method is that future expansion of new products and their locations in the operational database can be handled easily in the data warehouse tables.

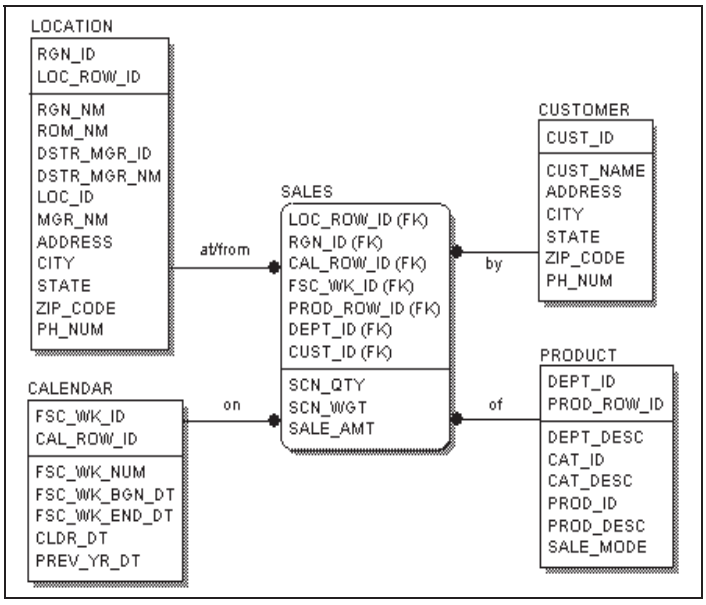


Figure 4. Method 4 - Row-wise distribution

Storing multi-values in the data warehouse star schema provides a more optimized structure than the normalized schema design of snowflakes. As shown in Figure 4, the sales, scan quantities and scan weight for each product and location can be stored separately. The disadvantage of this method is that it will have some duplicate values. More detailed advantages and disadvantages of the different methods will be discussed further in Section 7.

5 Software Design and Data Mining

The software used for design and implementation of the data warehouse system was Oracle and the interface was Java with embedded SQL using the Linux environment to calculate the data mining query response time and performance. For the data warehouse, the four described methods were created in the Oracle database system and the records were added to each database tables by making sure the dimensional tables supported similar collections of the record structures and numbers as other methods. The performance evaluation was carried out by running queries of various degrees of complication using Java to record the query run times. Three sets of queries were designed to provide each method with the same data mining and information retrieval criteria for the designed warehouse systems. These queries were considered as simple, complex and aggregated. The sample queries are listed below.

The performance for a simple query involved the core tables only, wherein the complexity of the snowflake design would not come into consideration. The following is the sample of simple query for the Method 1.

SELECT L.LOC_ID AS LOC_ID,P.PROD_ID AS PROD_ID, P.PROD_DESC AS PROD_DESC,C.CLDR_DT AS CLDR_DT,D.CUST_ID AS CUST_ID,D.CUST_NAME AS CUST_NAME,COUNT(1) AS ROW_COUNT **FROM** LOCATION L, PRODUCT P, CALENDAR C, CUSTOMER D, SALES S **WHERE** L.LOC_ID=S.LOC_ID AND P.PROD_ID=S.PROD_ID AND C.CLDR_DT=S.CLDR_DT AND D.CUST_ID=S.CUST_ID AND L.LOC_ID=353 AND C.CLDR_DT = '8-FEB-15' AND P.PROD_ID=785150 AND D.CUST_ID=3065789 GROUP BY L.LOC_ID, P.PROD_ID, P.PROD_DESC,C.CLDR_DT, .CUST_ID,D.CUST_NAME;

The complex query involves a "Where" condition for all the tables and the hierarchy. This associates all the joins and tables in the respective designs. The following is a sample of the complex query for Method 1.

SELECT L.LOC_ID as LOC_ID,L.DSTR_MGR_NM as DSTR_MGR_NM,L.ROM_NM as ROM_NM,P.PROD_ID as PROD_ID,P.PROD_DESC as PROD_DESC,P.CAT_DESC as CAT_DESC,P.DEPT_DESC as DEPT_DESC,C.FSC_WK_ID as FSC_WK_ID,D.CUST_ID as CUST_ID,D.CUST_NAME as CUST_NAME,D.CITY as CUST_CITY,SUM(S.SALE_AMT) as SALES_AMT,SUM(S.SCN_QTY) as SCN_QTY, SUM(S.SCN_WGT) as SCN_WGT **FROM** LOCATION L, PRODUCT P, CALENDAR C, CUSTOMER D, SALES S **WHERE** L.LOC_ID=S.LOC_ID AND P.PROD_ID=S.PROD_ID AND C.CLDR_DT=S.CLDR_DT AND D.CUST_ID=S.CUST_ID AND L.LOC_ID=2288 AND L.DSTR_MGR_NM='TONY CHANDLER' AND L.ROM_NM='EDUARDO GARCIA' AND C.CLDR_DT BETWEEN '05-FEB-15' AND '09-FEB-15' AND C.FSC_WK_ID=201506 AND P.PROD_ID=437124 AND P.CAT_DESC='CHICKEN' AND P.DEPT_DESC='MARKET' AND D.CUST_NAME='Peggy Nolen' GROUP BY L.LOC_ID, L.DSTR_MGR_NM, L.ROM_NM,P.PROD_ID, P.PROD_DESC, P.CAT_DESC, P.DEPT_DESC, C.FSC_WK_ID, D.CUST_ID,D.CUST_NAME,D.CITY;

An aggregated query is a step more complex query as it performs an aggregation operation along with involving all the joins and tables in the designs. The following is a sample of an aggregated query for Method 1.

SELECT L.RGN_NM as RGN_NM,L.ROM_NM as ROM_NM,P.DEPT_ID as DEPT_ID,P.DEPT_DESC as DEPT_DESC,C.FSC_WK_ID as FSC_WK_ID,C.FSC_WK_BGN_DT as FSC_WK_BGN_DT,C.FSC_WK_END_DT as FSC_WK_END_DT,SUM(S.SALE_AMT) as SALES_AMT,SUM (S.SCN_QTY) as SCN_QTY,SUM(S.SCN_WGT) as SCN_WGT **FROM** LOCATION L, PRODUCT P, CALENDAR C, CUSTOMER D, SALES S **WHERE** L.LOC_ID=S.LOC_ID AND P.PROD_ID=S.PROD_ID AND C.CLDR_DT=S.CLDR_DT AND D.CUST_ID=S.CUST_ID AND L.RGN_NM='Jacksonville' AND C.FSC_WK_ID=201506 AND P.DEPT_DESC='GROCERY' AND D.CITY='Jacksonville' GROUP BY L.RGN_NM, L.ROM_NM, P.DEPT_ID, P.DEPT_DESC, C.FSC_WK_ID, C.FSC_WK_BGN_DT, C.FSC_WK_END_DT;

The queries for the other methods were developed in the same manner to provide the same results. The time complexity for each method was tallied and tabulated.

6 Performance Analysis for Methods

For performance, each database was populated with the necessary records for queries to calculate the performance by the query response time. The records for each method were arranged to represent the same size and structure. Each query was processed 5 times and the average was calculated.

In the star schema, the records for each of the tables were sales (698), locations (5), products (5), customers (8) and calendar records (28). As shown in Figure 5, the result of processing for the simple query averages was 17.6 milliseconds. On running the complex "Where" clause query, an increased performance impact averaging 23.2 milliseconds was found. The aggregated query showed a lesser increase in query runtime averaging 23.8 milliseconds.

In the snowflake schema for the same records, the query run time for the simple query averages 18.2 milliseconds which was slightly more than in the star schema. On running the complex "Where" clause, the time difference increased further to 25.2 milliseconds, with the maximum difference seen while running the aggregated query at almost 3.8 milliseconds more than star schema.

For the extended star schema with column-wise in Method 3 and the extended star schema with row-wise in Method 4, there are lower processing times for the complex and aggregated queries as shown in Figure 5.

7 Technical Details of Methodologies

In this section, the different data warehousing methods are briefly described including advantages and disadvantages.

Method 1: Star Schema
In the star schema, the fact table is surrounded by four dimensions to form a star schema. The simplest model is also the most effective and preferred choice for design and implementation of data warehouses.

JOIN description for Method 1: In this design, the dimensions are joined to the fact table using a primary-foreign key relationship. For example, the location dimensional table and the sales fact table are joined at the lowest granular level LOC_ID.

Advantages of Method 1
1. The design of the star schema has an ease in implementation. At its simplest, the star schema is presented by a fact table surrounded by dimensions. Multiple star schemas are possible by the concept of composite dimensions which results in fact constellations.
2. Query performance is generally best for this design due to the fewer number of joins and tables within the query.

Disadvantages of Method 1
1. Redundancy is a drawback. Due to denormalization, a lot of duplicates are observed in the data. While this has become less of a drawback with hardware memory becoming cheaper, this can affect query performance when the size of the dimensions is very large.
2. Data loads are more time consuming to process because of the denormalized nature of the data.

Method 2: Snowflake Schema
A snowflake schema is more complex in the implementation as it normalizes the dimension tables. It resolves the issues associated with the star schema, where duplicated and multi-values are removed because of normalization and data load times are much faster. However, because of its complexity, the query response time is affected.

JOIN description for Method 2: In this design, the dimensional tables are normalized into more than 1 table and the dimension table with the lowest grain generally is joined to the fact table. The normalized dimension is connected to the tables at the various hierarchy levels. For example, the lowest granular location dimensional table LOCATION and the sales fact table are joined at the LOC_ID column level, while dimensional tables LOCATION is joined to the next granular level dimensional table LOCATION_1 at DSTR_MGR_ID level and similarly.

Advantages of Method 2
1. The redundancy problem associated with the star schema is resolved with the implementation of normalization.
2. Data load issues and loading time are much faster because of the high degree of normalization.

Disadvantages of Method 2
1. Query retrieval is impacted because of the increased number of tables and joins involved.
2. Because of the complexity of the design, maintenance becomes more complicated in the case of a large complex data warehouse.

Method 3: Extended Column-Wise Star Schema
A proposed alternate design is the extended column-wise star schema. In this design, the star distribution is preserved by expanding the replicating data in columns one after the other to maintain the star schema model. Thus, the star schema model is maintained. However, the design can be only be implemented when there are known number of multi-valued attributes, which requires the data modeler to have a very good knowledge of the number of multi-valued attributes for each column. A change in the fact table is required, as the fact table will only contain values at the highest aggregation across all dimensions, which reduces the data in the fact table significantly. The measures at the individual dimensional level are passed on to the distinct dimensions.

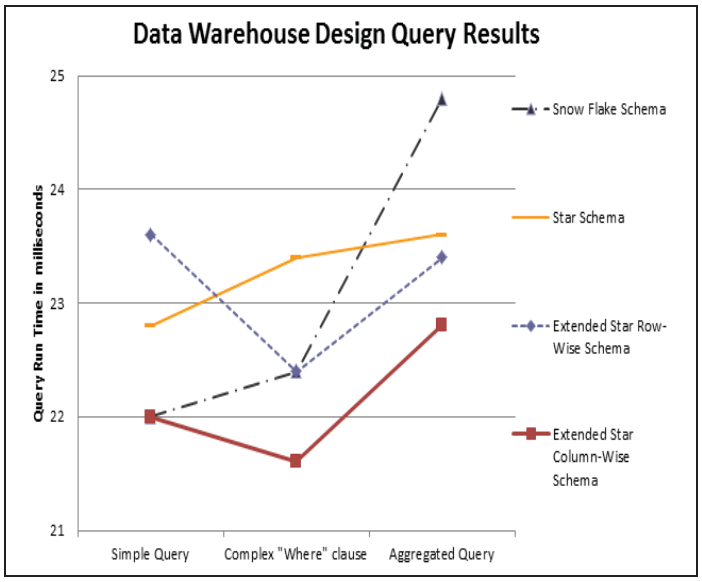


Figure 5. Data warehouse query result for Methods 1-4

JOIN description for Method 3: In this design, the joins between the dimensions and the fact table are at the highest grain. For example, the location dimensional table and the sales fact table are joined at the RGN_ID level.

Advantages of Method 3
1. Much faster query response as the star schema is maintained.
2. Overcomes the duplication issue associated with star schema without normalization.

Disadvantages of Method 3
1. The model works only when the number of multi-valued attributes for each column is known.
2. Data load is still an issue and can be error prone if not done correctly as per data model and requires data transformation.
3. A thorough knowledge of the data model is required placing a lot of reliance on data.
4. Data redundancy will occur as not all columns have data for multi-valued attributed columns.

Method 4: Extended Row-Wise Star Schema
A proposed second alternate design is the extended row-wise star schema. In this design, the star distribution is preserved by expanding the replicating data in rows vertically to maintain the star schema model. The dimensional keys start from the highest grain and with a corresponding row id as a composite key to define the relationship between the dimensions and fact. It can be utilized when the data modeler does not know the number of repeating attributes in each column. As the total number of attributes are unknown, for each column every new attribute is added row-wise by increasing the defined row count. A dimensional row id is added. This model also results in changes to the fact table, wherein the primary key are now comprised of all the highest grains dimensions and their row ids.

JOIN description for Method 4: The joins are not the simple join between a column in two tables, but a composite key comprising of a higher granular level (highest hierarchy) column and a dimensional row id. These result in a multi-column join between the dimensions and the fact table. The fact table keys are comprised of the higher granularity plus dimensional row ids as a composite foreign key. For example, the location dimensional table and the sales fact table are joined at the composite key level comprised of the RGN_ID and LOC_ROW_ID columns.

Advantages of Method 4
1. This design can be implemented when the number of multi-valued attributes is unknown.
2. Good query response times for complex queries.

Disadvantages of Method 4
1. Because of the composite key join, simpler queries may take longer to execute.
2. Data loads are still an issue as the data needs to be transformed to suit the row id – composite key implementation.
3. Data redundancy will occur as not all columns may have data for multi-row id attributes.

8 Conclusion

One of the most critical steps in the developmental process of a data warehouse is the provision of a design that will produce optimal functioning of the data retrieval needs for the Decision Support System. An important structure in the design of a data warehouse is the star schema which provides a design structure that allows a variety of queries to be efficiently processed. Based on the performance analysis and the results observed, the star schema is the best performing design for data warehousing. In data warehousing, where data can run up to petabytes, the ability to do complex analysis and fast query retrieval is generally desired. The star schema design is best suited to do so. However, the star schema has its own drawbacks which are not limited to duplication, high data load time and huge dimensional tables. These issues are solved by the snowflake design model, which normalizes the dimensions to overcome the drawbacks associated with the star schema. This normalization in turn introduces more complexity in the model in the form of new tables and joins, which in turn affects query performance. By maintaining the multiple values within the star structure, the data warehouse benefits from the preservation of the star schema and using a column-wise or row-wise design which results in efficiency of query processing.

The consideration of different methods in the design of a data warehouse, such as the ones proposed in this paper, will result in more optimized warehouses for the storage and retrieval of data through data mining by Decision Support Systems.

9 References

[1] Stella Gatziu, M. Jeusfeld, M. Staudt, and Y. Vassiliou. "Design and Management of Data Warehouses Report on the DMDW'99 Workshop", ACM SIGMOD Record, 28(4):7-10, December 1999.

[2] "Oracle Database Data Warehousing Guide", http://docs.oracle.com/cd/B19306_01/server.102/b14223/toc.htm, last accessed May 25, 2015.

[3] "Microsoft Modern Data Warehouse", http://www.microsoft.com/en-us/server-cloud/solutions/modern-data-warehouse/overview.aspx, last accessed May 25, 2015.

[4] W. H. Inmon. Building the Data Warehouse, John Wiley and Sons, 1996.

[5] Surajit Chaudhuri and D. Umeshwar. "An Overview of Data Warehousing and OLAP Technology ", ACM SIGMOD Record, 26(1):65-74, March 1997.

[6] Chuck Ballard, Daniel M. Farrell, Amit Gupta, Carlos Mazuela, and Stanislav Vohnik. "Dimensional Modeling: In a Business Intelligence Environment", IBM Redbooks, March 2006.

[7] Alfredo Cuzzocrea, Ladjel Bellatreche, Il-Yeol. "Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions", DOLAP '13: Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, October 2013.

[8] George Colliat. "OLAP, Relational, and Multidimensional Database Systems", ACM SIGMOD Record, 25(3):64-69, September 1996.

[9] Themistoklis Paplpanas. "Knowledge Discovery in Data Warehouses", ACM SIGMOD Record, 29(3):88-100, September 2000.

The Effect of Neighbor Selection in Collaborative Filtering Systems

S. Lee
Dept. of Computer Education, Gyeongin National Univ. of Education, Anyang, Korea

Abstract— *Collaborative filtering-based recommdender systems can aid online users to choose items of their preference by recommending items based on the preference history of other similar users. Similarity calculation plays a critical role in this type of systems, since the rating history of other users with higher similarity is given higher priority in recommendations. This study investigates qualifying conditions for choosing users to consult their rating histories. The conditions are addressed in two aspects: similarity between two users and the ratio of the number of items co-rated by two users. Through extensive experiments, thresholds yielding the best performance are obtained, thus the proposed strategy is found to achieve significant performance improvement.*

Keywords: Recommender system, Similarity measure, Memory-based collaborative filtering, Content-based filtering

1. Introduction

One of the popular ways to handle information overload in Internet is recommender systems. These systems offer users services that can help them to choose items of their preference. The services are basically made by referring to the history of item preferences of other customers. This method named *collaborative Filtering* (CF) has been popularly used in recent commercial systems to recommend items to online customers. Some of the successful CF systems are the Tapestry system, GroupLens, Video Recommender, Ringo, Amazon.com, the PHOAKS system, and the Jester system [9].

For the CF system to consult the history of items preferred by other users, it gives higher priority to those users more similar to the current user; this type is called *user-based* system. Hence, similarity calculation between two users is one of the most critical aspects of CF systems, which greatly influences recommendation performance. Most commonly used similarity measures are classified into correlation-based and vector cosine-based. Examples of the former approaches are the Pearson correlation and its variants, constrained Pearson correlation and Spearman rank correlation [1]. Cosine-based method treats each user as a vector and measures an angle between the vectors.

However, the volume of the history or the number of similar users along with their similarities are very important for successful recommendations. *Data sparsity* is a key issue that has been addressed by several researches. In particular, this problem is evident for a new user entered to the system, known as *new user problem* [2]. In this situation, it is almost infeasible to compute similarity or to be confident on the accuracy of the computed similarity [8]. To overcome this problem, some recent works has been proposed by utilizing the number of items commonly rated by two users [3], [5], [7], [8].

This paper investigates the effect of similarities and the number of co-rated items onto the recommendation performance. Further, it suggests a proper range of similarities and the ratio of the number of co-rated items yielding the best performance, through extensive experiments. The main contribution of the paper comes from its findings that when consulting ratings of other users for item recommendation, discarding those users dissatisfying given conditions rather than including them with low priorities as done by existing CF methods results in much better performance.

2. Proposed Approach

2.1 User-based Collaborative Filtering

The key assumption of this method is that if two users have similar tastes for items, their preference for new items would be also similar. From this assumption, the system can predict a rating of a new item for the current user, thus recommends those items with the highest predicted ratings. Similarity between two users is estimated based on the history of ratings made by the two users. The quality of the system is mostly dependent on whether the current user would like the recommended items with the highest predicted ratings.

Rating prediction for an item x for a current user u is made by following the steps below.

1) Select the nearest neighbors of user u who are most similar to u.
2) Let N_u be the nearest neighbors of user u who has rated item x.
3) Rating prediction for item x, $r'_{u,x}$, is made based on the Resnick's formula [6] as follows.

$$r'_{u,x} = \overline{r}_u + \frac{\sum_{v \in N_u} sim(u,v) \cdot (r_{v,x} - \overline{r}_v)}{\sum_{v \in N_u} |sim(u,v)|},$$

where $\overline{r}_u$ is the mean of the ratings made by user u and $sim(u,v)$ is the similarity value between users u and v.

As addressed above, the ratio of common history between two users is critical in calculating similarity. Hence, there have been some researches to take care of this issue which

mainly reflects the number of co-rated items into similarity calculation. Weighted Pearson correlation coefficient takes confidence on the neighbor into consideration, where the confidence increases with the number of common rated items [8]. Jamali and Ester [5] suggested a sigmoid function-based similarity measure which can produce low similarity if the users share the fewer common items. Rather than devising a new similarity metric, Bobadilla et al. [3] incorporated a factor reflecting the number of co-rated items into an existing similarity metric. The combined factor is the Jaccard measure [7], which is multiplied with the mean squared difference (MSD).

Jaccard is the metric that quantifies how many ratings are given to items in common by the two users [7]. Specifically, let I be the set of items and $r_{u,i}$ is the rating of item i given by user u. Then

$$I_u = \{i \in I \mid r_{u,i} \neq null\} \tag{1}$$

$$Jaccard_{u,v} = \frac{|I_u \cap I_v|}{|I_u| + |I_v| - |I_u \cap I_v|} \tag{2}$$

2.2 Preliminary Experiments

In order to investigate how different are the ratings between two users with different similarities, we conducted experiments on the two popular datasets, MovieLens and Jester [1]. The objective of these experiments is to figure out how similarity between two users is related with the difference of their rating behaviors. The characteristics of the dataset is presented in Table 1. It is expected that mean rating difference made by two users with higher similarities would be obviously lower. However, as discussed in [2], unusual behavior can sometimes be observed from some similarity measures for certain cases such as sparse datasets. Hence, it is worthwhile to compare the behaviors of rating differences between two users using different similarity measures.

Table 1: Descriptions of the datasets.

	MovieLens	Jester
Number of ratings	100,000	57,705
Rating matrix size	943×1,682	998×100
Rating scale	1 to 5 (integer)	−10 to +10 (real)
Sparsity Level	0.9370	0.4218

We calculated the mean rating difference of an item for varying similarities measured by Pearson correlation (PRS), cosine similarity (COS), and the mean squared difference (MSD). Figure 1 shows the result on the two datasets. As the similarity increases, MSD demonstrates the steepiest decrease in rating difference on both datasets, suggesting higher confidence on the measure than on the others. On the contrary, PRS seems to be the worst in terms of such confidence, since it shows the slowest decrease of rating difference with PRS similarity, implying that the performance gain achieved by consulting neighbors with higher similarity would not be that much, as compared to the case of MSD.

We also conducted experiments analogous to the above but with the Jaccard index. Instead of the mean rating difference, we focus on the ratio of each difference value per interval of Jaccard indices. As seen in Figure 2, the common behavior of both results on different datasets is that the ratios of higher difference are generally lower for the Jaccard index larger than some threshold. This behavior is not the case for low indices, because somewhat coarse outputs are obtained on the Jester dataset and the ratio of difference zero (diff=0) is found lower than that of difference one (diff=1) on the MovieLens. From these experimental findings, it is conjectured that excluding neighbors with low similarities or Jaccard indices might enhance the mean accuracy of predicted ratings, which motivated further experiments.

2.3 Experimental Results

In our experiments, 80% of each user's ratings were used for training, i.e., for calculation of similarity and Jaccard index, and prediction of ratings was made on the rest 20% of testing set. The accuracy of prediction can be measured through several metrics in literature [4]. In this paper, we use MAE (Mean Absolute Error) to estimate the prediction precision of the methods, which is defined as $MAE = \sum_i |r_i - r'_i|/N$, for N predictions r'_i for the corresponding real ratings r_i. Besides measuring prediction quality through MAE, we also evaluate the methods in terms of recommendation quality through precision and recall metrics [9]. We will use F1 in our evaluations, as it is a comprehensive indicator combining both precision and recall with equal weights.

The baseline similarity metrics of our experiments are selected from most commonly used ones in CF systems, i.e., Pearson correlation (PRS), the cosine similarity (COS), and the mean squared difference (MSD). From the discussion in the previous subsection, selection of neighbors are made satisfying conditions of similarity and Jaccard index less than some threshold. We performed extensive experiments using different thresholds to obtain the best ones. The resulting measures are named PRS_CMB, COS_CMB, and MSD_CMB, for the corresponding PRS, COS, and MSD, respectively.

Figure 3 shows the results on the MovieLens dataset performed for varying number of nearest neighbors (top NN). It is notable that simply excluding those with low similarities or Jaccard indices can greatly improve prediction accuracy as well as recommendation quality. This is especially the case with PRS and MSD. Such improvement becomes greater along with top NN, implying that consulting ratings of neighbors with low similarities or Jaccard indices affects performance negatively, even though their ratings are reflected with low weights according to the Resnick's formula [6].

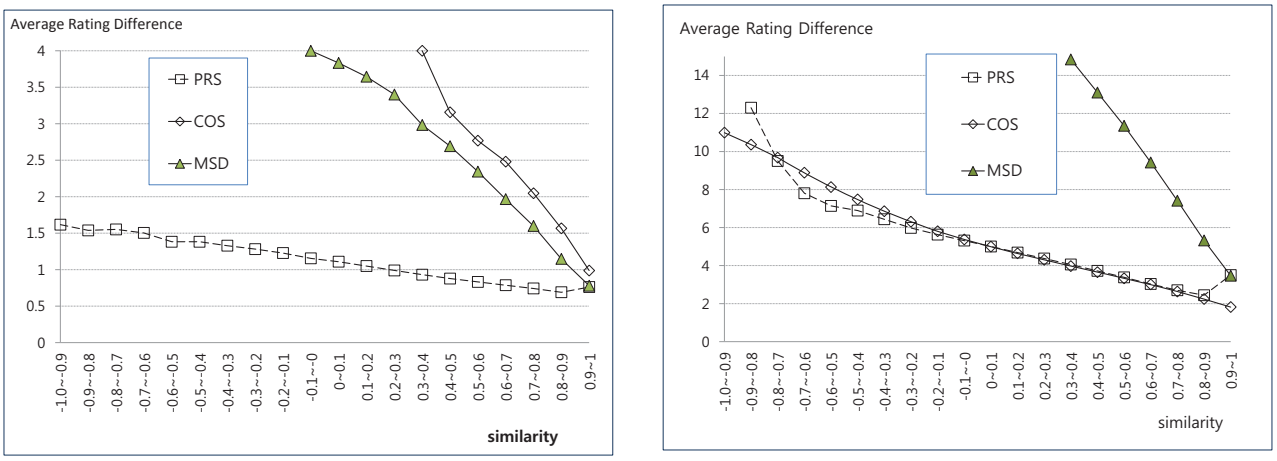


Fig. 1: Average rating difference of an item between two users with different similarities: MovieLens (left) and Jester datasets (right).

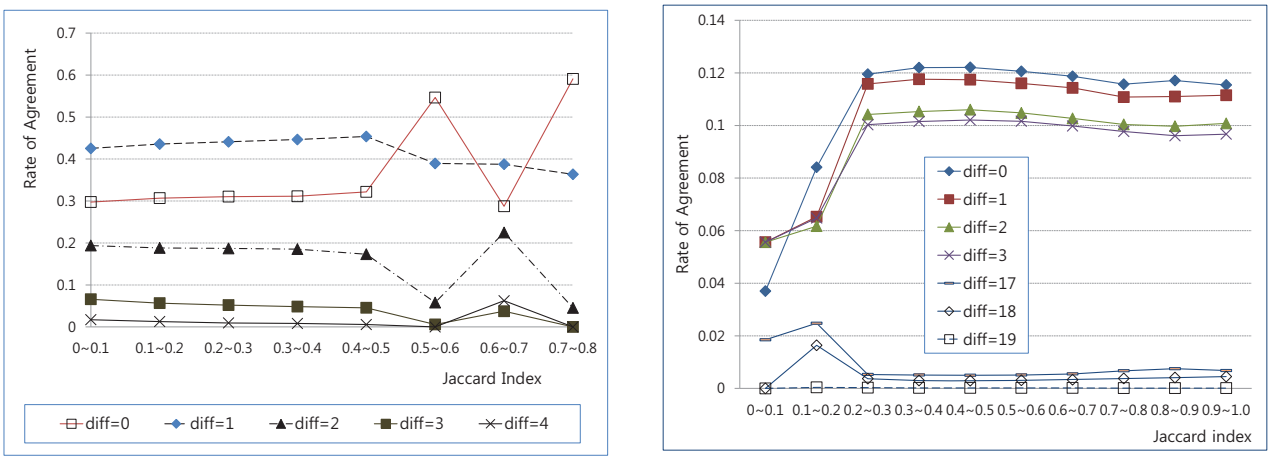


Fig. 2: Ratio of each rating difference per interval of Jaccard indices: MovieLens (left) and Jester datasets (right).

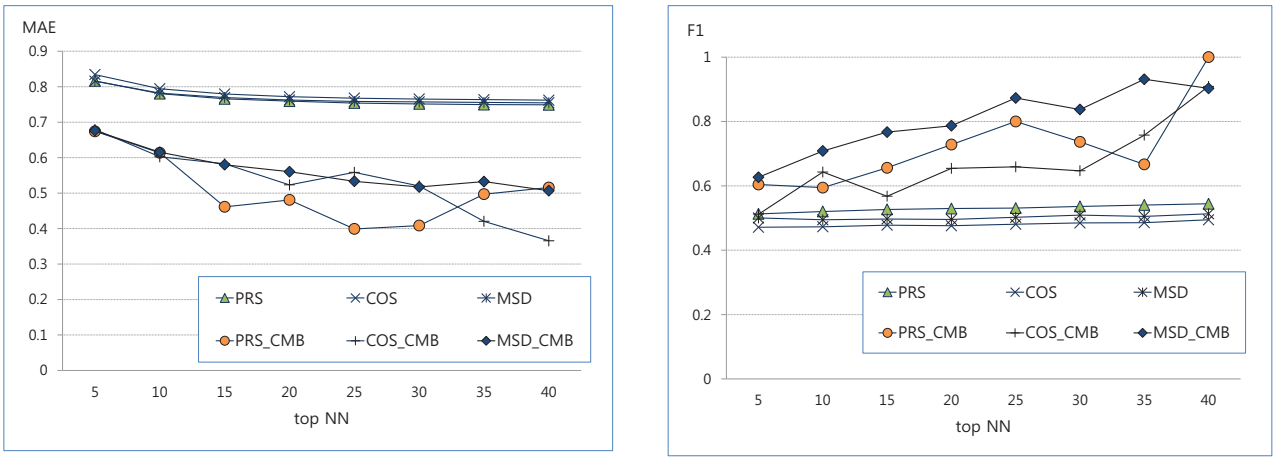


Fig. 3: MAE and F1 performance on the MovieLens dataset.

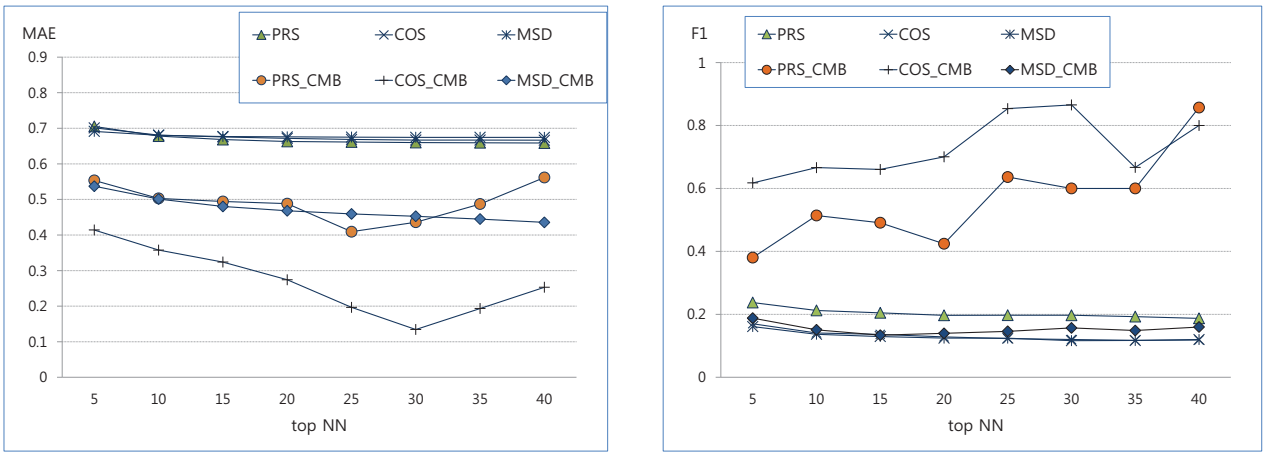


Fig. 4: MAE and F1 performance on the Jester dataset.

Our next experiments used the Jester dataset to evaluate the methods. In Figure 4 we can observe similar trend as in the MovieLens dataset, but notably higher improvement. The reason for such higher improvement is partly due to the lower sparsity level of the Jester dataset, where there would still remain neighbors to consult ratings even after exclusion of those with low similarities or Jaccard indices. Moreover, there is higher probability that these remaining neighbors would have high similarities or Jaccard indices than in the case of the MovieLens dataset, thus yielding higher accuracy of predicted ratings. Especially, COS benefits a lot more from the cautious selection of neighbors imposed by our strategy. The reason might be that the ratio of rating difference of zero increases more sharply with COS than with PRS or MSD in our experiments, thus gives higher accuracy when consulting neighbors of higher similarities, details of which are not presented here due to the space constraints.

3. Conclusions

This paper demonstrates that simple constraints for selecting neighbors to consult their ratings yield much better performance in collaborative filtering systems. The selection criteria are made based on two indices, similarity and Jaccard index. Extensive experiments are conducted to search for the thresholds of indices yielding the best performance. The results indicate that the proposed strategy outperforms the baseline measures in both aspects, prediction quality and recommendation quality. Furthermore, it is found that the strategy is in general more effective on the denser dataset having the larger rating scale.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowledge & Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[2] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Information Sciences*, vol. 178, no. 1, pp. 37–51, 2008.

[3] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, "A collaborative filtering approach to mitigate the new user cold start problem," *Knowledge-Based Systems*, vol. 26, pp. 225–238, 2011.

[4] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Information Systems*, vol. 22, no. 1, pp. 5–53, 2004.

[5] M. Jamali and M. Ester, "TrustWalker: A random walk model for combining trust-based and item-based recommendation," in *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 397–406.

[6] B. Jeong, J. Lee, and H. Cho, "Improving memory-based collaborative filtering via similarity updating and prediction modulation," *Information Sciences*, vol. 180, no.-5, pp. 602–612, 2010.

[7] G. Koutrica, B. Bercovitz, and H. Garcia, "FlexRecs: Expresing and combining flexible recommendations," *SIGMOD*, 2009, pp. 745–757.

[8] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, "A new user similarity model to improve the accuracy of collaborative filtering," *Knowledge-Based Systems*, vol. 56, pp. 156–166, 2014.

[9] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009.

Multidimensional Analysis Using Knowledge Layer for Effective Decision Making

Ayaz Mahmood, Dr. Sohail Asghar
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST),
Islamabad, Pakistan.
Department of Computer Sciences
COMSATS Institute of Information Technology, Islamabad, Pakistan.
ayaztg@gmail.com sohail.asg@gmail.com

Abstract

Business intelligence is one of the most focusing areas for researchers in the last few years for providing various solutions using the concept of Knowledge Discovery. Various researchers have proposed their own models to utilize knowledge discovery for efficient use of information to enable enhanced and effective decision making process. The paper is supplemented with validation of the proposed framework through a case study. Further, the proposed model is visualized through a flowchart for better understanding of the process model. In addition, implementation details are also discussed. Finally, a dashboard is built on the top for the proof of concept.

Keywords: Business Intelligence, Knowledge Discovery, Decision Support System, Data Mining.

1. Introduction

Business Intelligence (BI) has become widely used technology due to rapid and massively increased volumes of data. Organizations are tending to utilize BI as a core source of information for decision making process. Modern BI tools operate on the data layer and decision making is done by different analysis reports. When a new query arises, finding associations between data, for instance, conventional BI tools may not be able to answer such queries. Using massive amount of data in a traditional BI system makes is harder to find hidden patterns in the data.

In order to improve effective decision making using BI systems, we transform data into valuable information and knowledge management practices are executed to handle such information to support decision making. In this paper we introduce the concept of Knowledge Layer to conventional BI system to provide the ability of knowledge-driven decision making. We have implemented Apriori algorithm at Knowledge Layer for generating association rules from database transactions. These rules reveal the hidden patterns and relationship among data sets of database transactions.

In data mining, Apriori algorithm is used for finding relationships between data sets of a transaction. These relationships are known as Association Rules. Support and Confidence are two key concepts in algorithm for searching relationships. Support is describes as the percentage of transactions in a data set which contains the item set. Whereas, Confidence is defined as $conf(X \rightarrow Y) = supp(X \cup Y) / supp(X)$.

With the assistance of Apriori algorithm implementation and association rules generation, ability of effective decision making of BI systems can be improved while answering new questions.

Apriori algorithm is devised to work on databases which contain data in the form of transactions. The aim of Apriori algorithm is to search relationships and associations between data sets of a database. Usually it is known as “Market Basket Analysis”. Each data set contains item set and is known as a transaction.

The algorithm produces sets of rules which shows how often items exists in a set of data which has gone through under algorithm. Following table illustrates the

concept in a better way. Each line refers to set of items and is called transaction.

Bread	Milk	Butter
Bread	Milk	Cheese
Bread	Milk	Egg
Bread	Milk	Cheese

1. 100% of sets with Bread also contain Milk
2. 25% of sets with Bread, Milk also have Butter
3. 50% of sets with Bread, Milk also have Cheese

Apriori algorithm utilizes breadth-first search and a Hash tree to calculate item sets from a transaction. Algorithm Pseudo code of Apriori for a database transaction **T,** and support € is given as follows. Where **Ck** is the set of candidates for level **k**. **count[c]** approaches a member of the data structure that corresponds to candidate set **c,** which is in the beginning taken as zero [3]. Diagram below shows the Pseudo code of Apriori algorithm:

```
Apriori(T, ε)
    L1 ← { large 1-itemsets }
    k ← 2
    while L_{k-1} ≠ ∅
        C_k ← {c | c = a ∪ {b} ∧ a ∈ L_{k-1} ∧ b ∈ ⋃L_{k-1} ∧ b ∉ a}
        for transactions t ∈ T
            C_t ← {c | c ∈ C_k ∧ c ⊆ t}
            for candidates c ∈ C_t
                count[c] ← count[c] + 1
        L_k ← {c | c ∈ C_k ∧ count[c] ≥ ε}
        k ← k + 1
    return ⋃_k L_k
```

Figure1. Pseudo code of Apriori algorithm [13]

In data mining, Association rule is a famous technique to discover hidden relationships between the data sets of transactions of a database or data warehouse. For instance, the rule {Bread, Milk} → {Egg} found in the sales data of a shop would show that if a customer buys Bread and Milk together, he or she is likely to buy Eggs as well [3]. This type of information can be precious for the decision making authorities for taking decisions about marketing strategies, e.g. product promotional pricing or placement of products on the shelf of stores.

If we formalize the problem (of finding associations between data sets of a transaction) then:

- Database Transaction T: set of target transactions. T = {t1, t2, t3… tn}
- Each transaction owns set of items I (item set)
- An item set is a combination or collection of items, I = {i1, i2, i3… im}

X → Y, is the form of putting relationships as an association rule. Following example makes the concept more clearly about association rules and related concepts.

TID	Items
T1	Bread, jelly, peanut-butter
T2	Bread, peanut-butter
T3	Bread, milk, peanut-butter
T4	Cheese, bread
T5	Cheese, milk

Bread → Peanut-Butter
Cheese → Bread

Following are the item sets that appear frequently together.

I = {Bread, peanut-butter}
I = {Cheese, bread}

Support Count (**σ**) is the Frequency of occurrence if an item set [14].
σ ({Bread, peanut-butter}**)** = 3
σ ({Cheese, bread}) = 1

Support is the ration of database transactions that owns item sets.
S ({Bread, peanut-butter}**)** = 3/5
S ({Cheese, bread}) = 1/5

The two most used concepts in association rules are:
Support (S) is the occurring frequency of rule i.e. num. of transactions containing both X and Y [14].
S = σ (X U Y) / # of transactions
Confidence (C) is the strength of association which is the measure of Y's occurrence in a transaction that contains X [14].
C = σ (X U Y) / σ (X)

Hence the support and confidence of the transactions mentioned in the above table are depicted in the following table.

TID	S	C
Bread → peanut-butter	0.60	0.75
peanut-butter → Bread	0.60	1.00
Cheese → bread	0.20	0.50
peanut-butter → jelly	0.20	0.33
jelly → peanut-butter	0.20	1.00
jelly → milk	0.00	0.00

The paper is organized as follows: This section contains introduction. Section II provides an overview of existing Knowledge management techniques, methods and frameworks. Proposed Knowledge Layer based model is provided in section III and section IV presents validation of propose framework through a case study. Implementation and Results interpretation is provided in Section V followed by conclusion and possible dimensions of future research work.

2. Literature Review

A number of different techniques and models have been proposed in the past to improve effective and efficient decision making via business intelligence tools.

Yong Feng [1] introduced multi-agent technology in BI systems for efficient decision making. The approach also proposed a low-cost BI framework for the analysis of core components' function and operation mechanism of the BI system. Proposed framework is based upon the layered architecture which consists of three layers mainly. Data resource layer is the backend layer, which handles all the tasks related to data ranging from data availability to data monitoring. Middle layer is known as core function layer, which consists of agent programs to perform functions like handling requests from users, process it and send results back to user. User management agent and OLAP agent are the main programs which performs respective functions. Front end layer is known as UI-Layer, which deals with the data representation to the prospective users or clients. Proposed solution is effective for low-cost BI system; however, it lacks the systematic implementation of the proposed framework. Authors only discussed the framework in detail, no implementation has carried out to support the idea presented in the paper.

Qing-sheng [2] proposed a Rosetta Net framework for business intelligence systems. Framework helps in realization of optimize business process. It also provides a cost effective BI solution. Proposed architecture claims seamless integration BI systems based upon Rosetta Net Implementation framework (RNIF). RNIF mainly has been developed with respect to Network environment as it deals with network protocols like HTTP, SMTP, TCP/IP and other transfer protocols. RNIF's core component is RosettaNet Business Message Component. The component mainly consist of two parts, one is Header, which contains metadata information attached to the actual message. Second is Service Content which contains the actual message. RosettaNet utilizes XML as a data source for the information exchange. For analytic purposes, RNIF utilizes Workflow messaging to interact with different business processes and workflows. The proposed approach is good to work with business workflows and processes, also with the integration of BI systems. Though, the proposed model needs to incorporate variety of data sources other than XML data source. This will enhance the flexibility and scalability of the proposed approach.

SHAN Wei [3] proposed knowledge mining model based upon Shannon Information Theory and Bayesian

estimation algorithm for generating association rules. The paper also discusses the knowledge evaluation system of the BI system. It helps in solving the data consistency and redundancy problem during the data acquisition stage in a BI system. Overall paper discussed knowledge mining in a BI system, solution of data quality problems were proposed by introducing the use of Bayesian estimation. Although, paper discusses the conceptual model and development, however, any concrete implementation on a real world scenario is missing to support the proposed architecture of knowledge mining technique in a BI system.

Tong Gang [4] introduced the idea of business process and knowledge management to the traditional BI system. The approach utilizes the implementation of case-based and rules-based reasoning technology. It also provides a knowledge management approach in BI systems.

Luan Ou [5] proposed a customized model of BI system for the retail industry. Proposed framework utilizes JAVA platform to develop custom applications based upon the different data sources. A custom software application for BI system for decision making has been developed for the proof of concept purpose. This system helps in profitability analysis, outside environment analysis, and KPI management, etc. The methodology is good for small scale applications but not feasible for big data enterprises.

Li Dalin [6] discusses design and development of a multi agent based BI system (MABBI). Proposed approach utilizes Database, Data Warehouse and Knowledge Discovery / Data Mining modules as main components of the model. MABBI uses agent technology to deal problems in the BI system.

Alexander Loebbert [7] discusses role of knowledge discovery in a web environment for BI system of retail business. It also proposed a model for web data mining based upon Euclidean distance, Mahalanobis distance and City block distance. Classification model technique is used in the approach to distribute data to different groups.

SHI Changqiong [8] proposed an integration framework for BI systems based on Ontology. In this research, a service-resource based ontology is proposed which includes Ontology mode and Data model patterns.

Xu Xi [9] proposed a BI system model based on Web-Services. The design of the overall framework is divided into three parts which are BI processing system (based on Web Services), commercial service distributors and BI services registration centre which are the end users.

A limitation observed during literature review is that these techniques proposed models which usually are based on traditional BI system architecture. These systems can answer the adhoc queries, manages KPI, provides information, etc but are unable to find the hidden patterns and relationships among data sets of the transactions. For instance, a user wants to see the relationship between items sold over the period of time. There is no such methodology proposed earlier for answering such queries.

3. Knowledge Layer Model

To handle aforementioned problem, we propose a multi-dimensional decision making model using knowledge layer. The proposed model is divided into four main layers as follow:

1. Data Layer
2. Knowledge Layer
3. Transformation Layer
4. Presentation Layer

Data Layer (DL): This layer directly deals with the data. This layer contains heterogeneous data sources such as flat files, xml data files, spreadsheets, enterprise resource planning system, or multiple operational databases such as MY SQL, Oracle, MS SQL Server. These components may serve as a standalone data source or they can be combination of hybrid data sources. Moreover, a single data ware house (DWH) could also serve as a data source. We assumed that in case of DWH as a data source, target data which needs to be analyzed, is already uploaded into DWH from operational data sources as mentioned earlier. The first step of our framework is data provisioning that is done by DL. DL supports and manages data extraction from the data sources and passes it to the upper layer.

Knowledge Layer (KL): Deals with the Knowledge management in terms of finding association and relationships between data sets provided by the DL. KL controls the execution of Apriori Algorithm and generates association rules. Target data which is extracted and provided by DL gets passed to KL. These transactions are then input into Apriori algorithm for

data mining. Apriori algorithm then generates association rules based upon the supplied transactions. Moreover, Apriori needs minimum 'Support' and 'Confidence' to be entered as Inputs for generation of association rules in the algorithm. Based upon these inputs Apriori produces association rules. Resultant rules contain the relationships between item sets of transactions. Later, these rules serve as input to the next layer

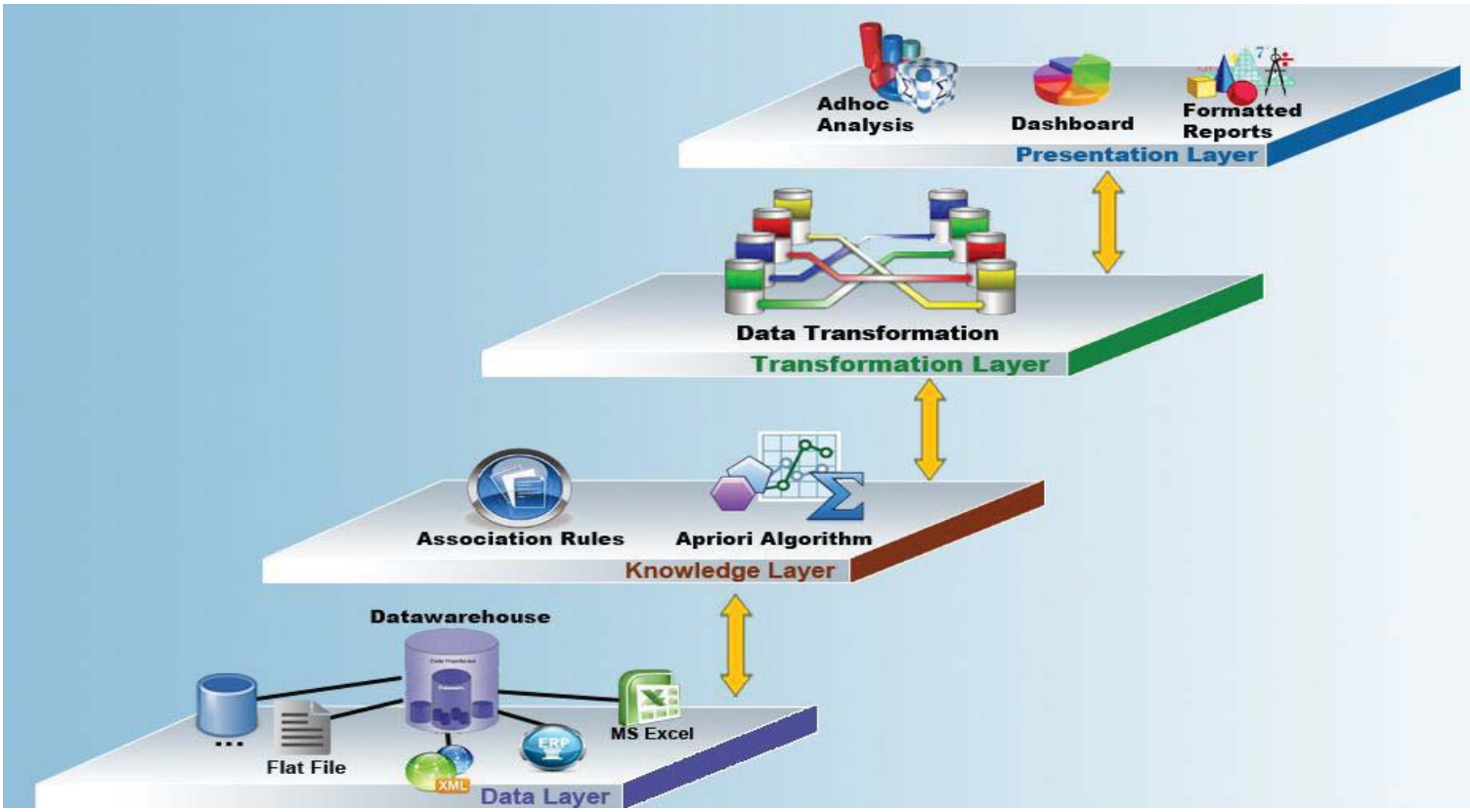
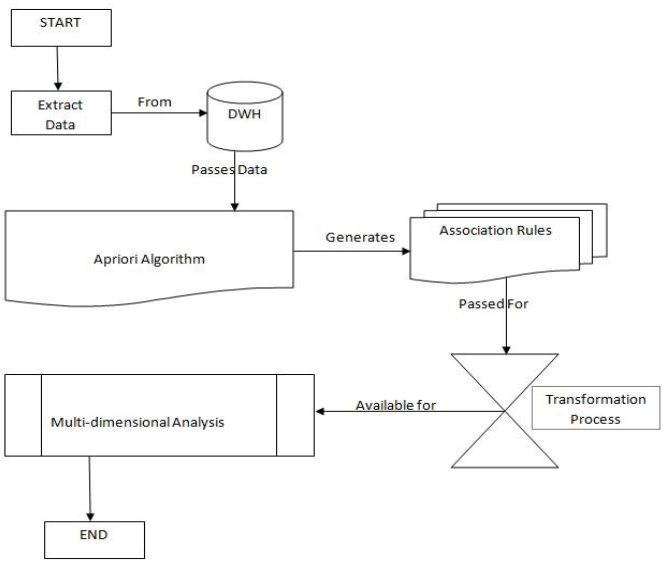


Figure 2: Framework for multidimensional analysis using Knowledge Layer.

Transformation Layer (TL): Supports and manages the data transformations. Since association rules in KL are in a specific format so they need to be transformed into required format for the analysis. For instance, {C→L = 100%} is an association rule. The output rule generated by Apriori algorithm is not clear enough to get the actual interpretation. Hence, to comprehend the rules, a utility is required to interpret association rules. We have implemented a utility program in JAVA which takes the association rules as input and interprets them accordingly based upon transformation logics. Therefore, the rule {C→L = 100%} gets transformed into {Cotton →Lawn = 100%} which is easier to understand.

Presentation Layer (PL): Displays required information to the users via OLAP tools such as adhoc analysis, dashboards, formatted reports etc. PL gets transformed rules from TL. Transformed rules are processed on this layer for presentation. This layer only



deals with displaying information to user. Later, user can perform multi-dimensional analysis upon the processed target data for effective decision making.

Figure 3 Flow Chart of process for proposed model.

Figure shows the process flow of the proposed framework. It starts with the extraction of data from DWH. Data extraction is performed on Extraction, Transformation and Loading (ETL) process. For the ease of understanding we name it target data. Once extracted, target data is passed to Apriori algorithm. Two input parameters i.e., Support and Confidence are also passed. Algorithms get executed and process the transactions from target data and generate association rules. The generated rules are in a semi formal format. These rules then undergoes to a transformation process for converting them into required format. After transformation, these rules get available for OLAP tools where user can perform multi-dimensional analysis which helps in effective decision making.

4. Case Study

We apply our proposed model on a case study based on SAP Business One (B1) ERP.

Problem Statement: A company Leisure Club (LC) is maintaining its daily transactional data in SAP BI – an ERP product by SAP. LC is a leading clothing brand in Pakistan. LC has implemented the SAP Business Warehouse (BW) module of SAP ERP for maintaining its DWH. They have implemented SAP Business Object (BO) to meet Business Intelligence (BI) demands. The top management of LC is interested to know the relationship between different cloth types during their sales of past two years, in order to develop better marketing and production strategies. Unfortunately, current OLAP analysis implementation is unable to answer such queries. We have implemented the proposed model in different tasks as follow:

Task 1: Data Extraction. Based upon the requirements of senior management we have extracted sales data for past two years that is from 2009 – 2011. Data gets extracted from the SAP BW using ETL process and finally loaded in to a text file. Text file contains sales data transactions for the required period of time. Regarding data to be processed, we took the different cloth types such as Chiffon, Silk, Cotton, Malmal, Swiss Lawn, Twill and Organza etc.

Task 2: Apriori Algorithm Implementation. We have implemented Apriori algorithm using JAVA programming language. It takes the transactions from the text file and processes them for the rules generation. We pass on transactions to the interface of program and then enter two input parameters i.e. Support and Confidence. Apriori Implementation executes the algorithm and show results in form of association rules. Our implementation takes items of the transactions as alphabets, for instance, "C" represents Cotton, "c" shows Chiffon, "T" points at Twill, "L" for Lawn and so on.

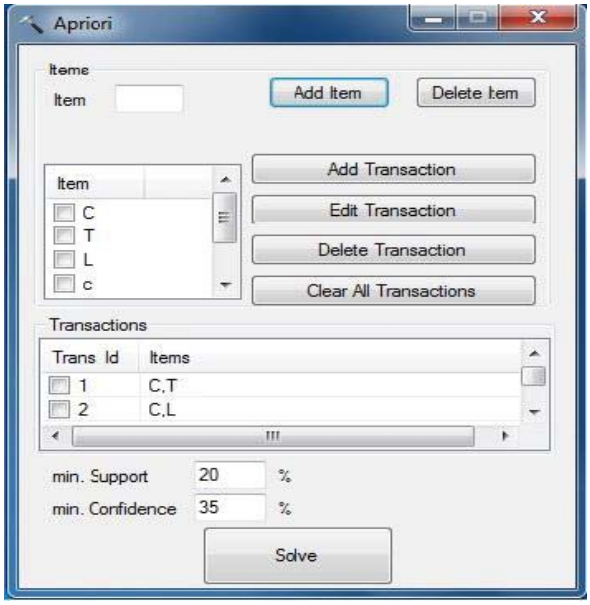


Figure 4 Prototype interface of Apriori algo.

We took 400 transactions from the text file and input them in interface. Interface allows add, edit and delete transactions if entered by mistake. Similarly, items of the transactions can also be add or delete from interface. After inserting the items and transactions, support and confidence parameters need to be filled and then press solve for association rules generation.

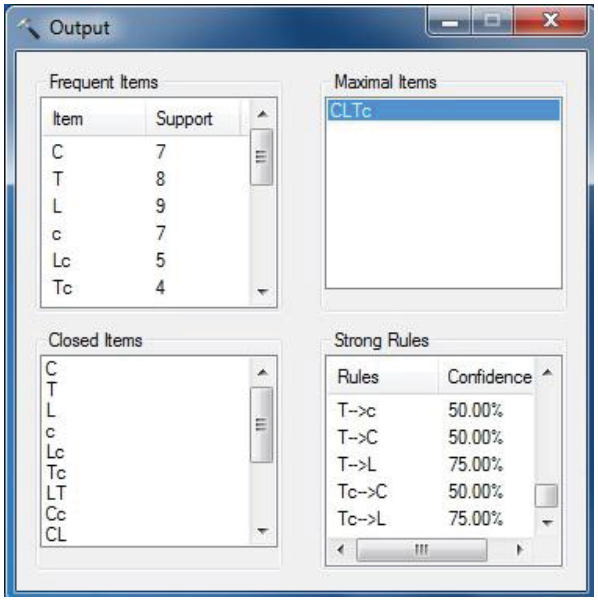


Figure 5 Association rules generated on the basis of data entered in to the interface.

Output interface shows Frequent Items and Strong Rules based upon the support and confidence entered previously. "Frequent Items" calculates and shows the support of items and their combination whereas, "Strong Rules" shows Association between items and their confidence. We got different association rules as a result such as {T→C = 50%}, {T→c = 50%}, {T→L = 75%} and so forth. These rules are generated on relationship between the items of each transaction. Ideally we are in search of the rules whose confidence is 100% but for the multi-dimensional analysis purposes we need all these rules to be displayed at OLAP analysis.

Task 3: Rules Transformations. Rules generated by Apriori program are in semi formal format. To convert them into required format for OLAP analysis tools, we have implemented a utility program in JAVA programming language. This program converts each association rule into the required format. For instance, {T→C = 50%} gets converted into {Twill → Cotton = 50%}, {L → C = 100%} becomes {Lawn → Cotton = 100%} and so on. This converted form is now easy to understand for analysis tools at presentation layer.

Task 4: Multidimensional Analysis. Finally, the transformed rules are sent to presentation layer for analysis purposes. User can perform analysis in many ways; such as, adhoc analysis, dashboards, and formatted reports. In our case study, we utilized the SAP BO platform for the analysis purpose. SAP BO is a combination of different tools for multidimensional analysis and reporting. Under the SAP BO platform, adhoc analysis is done by SAP Web Intelligence (Webi) tool, dashboards are designed in SAP Dashboard Designer tools and formatted reports are created in SAP Crystal Reports.

We have implemented the results in SAP dashboard designer tool by creating a dashboard on top of transformed rules. Dashboard allows users to select and evaluate different cloth types and analyze their confidence. For instance, user can select cotton and silk and see their association during the last two years. Based upon this analysis user can do effective decision making.

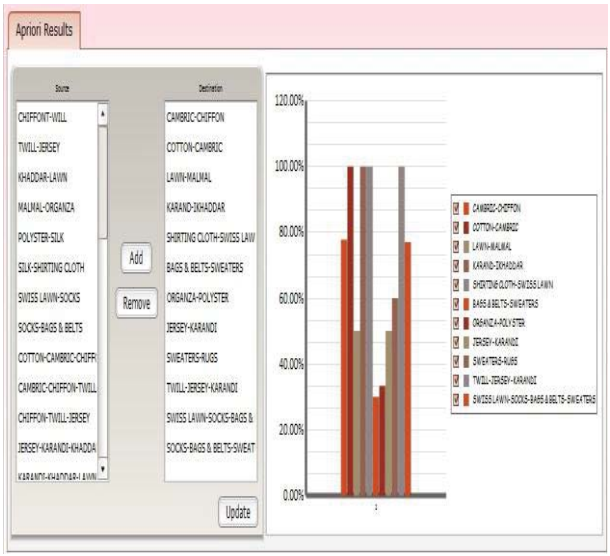


Figure 6 the dashboard built on top of transformed rules.

Interpretation

Dashboard is used to interpret and evaluate the results of association rules. For better understanding of analysis the rules needs to be interpreted. For Instance, {Twill → Cotton = 50%} shows that chances of selling twill with cotton were 50% during the last two years according to processed transactions. Similarly, {L → C = 100%} depicts that chances of lawn to be sold in combination with cotton were 100%. Now while looking at these two rules, management can decide that there is a need to produce more lawn and cotton (Production Policy) and place them on selling shelf consecutively (Marketing Strategy). Apparently these are just some processed rules; however, these rules can deduce new policies and strategies as explained above. In this way, senior management can do effective decision making from multidimensional analysis using knowledge layer.

Implementation

As mentioned earlier, KL is responsible for implementing Apriori algorithm and we have implemented it in JAVA. Apriori Implementation consists of:

Class AprioriCalculation: Generates Apriori item sets from transactions. Class consists of following methods:

aprioriProcess (): used to generate the Apriori item sets.

```
do
        {
            //increase the itemset that is being looked at
            itemsetNumber++;

            //generate the candidates
            generateCandidates(itemsetNumber);

            //determine and display frequent itemsets
            calculateFrequentItemsets(itemsetNumber);
            if(candidates.size()!=0)
            {
                System.out.println("Frequent " + itemsetNumber +
"-itemsets");
                System.out.println(candidates);
            }
        //if there are <=1 frequent items, then its the end.
This prevents reading through the database again. When there is
only one frequent itemset.
        }while(candidates.size()>1);
```

generateCandidates (): Generate all possible candidates for the n-th item sets

```
private void generateCandidates(int n)
    {
        Vector<String> tempCandidates = new Vector<String>();
        String str1, str2;
        StringTokenizer st1, st2;
        if(n==1)
        {
            for(int i=1; i<=numItems; i++)
            {
                tempCandidates.add(Integer.toString(i));
            }
        }
        else if(n==2) //second itemset is just all combinations
of itemset 1
        {
            //add each itemset from the previous frequent
itemsets together
            for(int i=0; i<candidates.size(); i++)
            {
                st1 = new StringTokenizer(candidates.get(i));
                str1 = st1.nextToken();
                for(int j=i+1; j<candidates.size(); j++)
                {
                    st2 = new StringTokenizer
(candidates.elementAt(j));
                    str2 = st2.nextToken();
                    tempCandidates.add(str1 + " " + str2);
                }
            }
        }
        else
        {
        candidates.clear();
        candidates = new Vector<String>(tempCandidates);
        tempCandidates.clear();
    }
```

Transformation Utility

This program is used to transform association rules into required format

```
if (itemset.contains('c')){
        newItem = Replace(item).equalto("Chiffon");
}
elseif (itemset.contains('C')){
        newItem = Replace(item).equalto("Cotton");
}
elseif (itemset.contains('L')){
        newItem = Replace(item).equalto("Lawn");
}
elseif (itemset.contains('S')){
        newItem = Replace(item).equalto("Silk");
}
```

Conclusion and Future Work

Our proposed model can be effective for multidimensional analysis which supports improved decision making to the decision authorities in an organization. Utilizing knowledge layer with the traditional BI architecture helps in discovering the hidden but useful information that lies within the data in a database or DWH. Thus, our proposed framework is capable enough to deliver the real insights of data which is difficult to capture with conventional architecture of BI systems.

Our proposed model provides effective decision making using knowledge layer. However, proposed approach is limited only to find relationships i.e.

generating association rules from data. It needs to incorporate other data mining or knowledge discovery techniques as well.

In future, we intend to extend and enhance the proposed architecture to include Clustering and Classification techniques of data mining for better decision making and knowledge discovery.

REFERENCES

1. Yong Feng, Yang Liu, Xue-xin Li, Chuang Gao and Hong-yan Xu, "Design of the Low-cost Business Intelligence System Based on Multi-agent," International Conference of Information Science and Management Engineering, 2010.
2. Qing-sheng Xie and Gui-xian Zhou, "Developing a Framework for Business Intelligence Systems Based on RosettaNet Frame," International Conference on Wireless Communications, Networking and Mobile Computing, 2008.
3. SHAN Wei and ZHANG Qing-pu, "Study on Knowledge Mining of the Business Intelligence System," International Conference on Wireless Communications, Networking and Mobile Computing, 2008.
4. Tong Gang, Cui Kai, and Song Bei, "The Research & Application of Business Intelligence System in Retail Industry," Proceedings of the IEEE International Conference on Automation and Logistics Qingdao, China September 2008.
5. Luan Ou and Hong Peng, "Knowledge and Process Based Decision Support in Business Intelligence System," Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06), 2006.
6. LI DALIN, "THE RESEARCH OF BUILDING BUSINESS INTELLIGENCE SYSTEM OF THE RETAIL BUSINESS IN WEB ENVIRONMENT," INTERNATIONAL CONFERENCE ON E-PRODUCT E-SERVICE AND E-ENTERTAINMENT (ICEEE), 2010.
7. ALEXANDER LOEBBERT AND GAVIN FINNIE, "A MULTI-AGENT FRAMEWORK FOR DISTRIBUTED BUSINESS INTELLIGENCE SYSTEMS," 45TH HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCE (HICSS), 2012.
8. SHI Changqiong and ZHANG Dafang, "Web Service-Based Business Intelligence System Research and Implementation," 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08), 2008.
9. XU XI AND XU HONGFENG, "DEVELOPING A FRAMEWORK FOR BUSINESS INTELLIGENCE SYSTEMS INTEGRATION BASED ON ONTOLOGY," INTERNATIONAL CONFERENCE ON NETWORKING AND DIGITAL SOCIETY, ICNDS '09, 2009.
10. B Azvine, Z Cui, DD Nauck and B Majeed, "Real Time Business Intelligence for the Adaptive Enterprise," Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06), 2006.
11. ZHIJUN REN, "BUILDING BUSINESS INTELLIGENCE APPLICATION WITH SAP BI," INTERNATIONAL CONFERENCE ON MANAGEMENT AND SERVICE SCIENCE, MASS '09, 2009.
12. R Abdullah and M Talib, "A model of knowledge management system for facilitating knowledge as a service (KaaS) in cloud computing environment," International Conference on Research and Innovation in Information Systems (ICRIIS), 2011.
13. http://www.en.wikipedia.org/wiki/Apriori_algorithm

Identifying Common Individual Identities across Various Social Media Sources through Analytics

Gurpreet Singh Bawa
Gurgaon, India

ABSTRACT:

In this work we also examine 5 different approaches for determining user similarity as reflected by activity in social media applications such as Facebook, Twitter etc.. Lists of similar people returned by the five approaches vary from each other as well as from the list of people the user is familiar with, so we also suggest an aggregation of approaches to produce an overall probability that each combination of user is same. An evaluation of the approaches and the overall aggregate demonstrate the utility across different scenarios, such as information discovery and expertise location, and also highlights sources and aggregates that are particularly valuable for inferring user similarity.

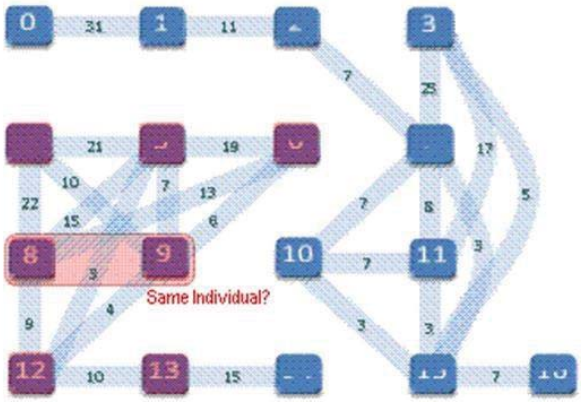
1. SIMILARITY APPROACHES

In order to understand the characteristics of different similarity approaches, several social media applications were analyzed (listed here in order of deployment): sample of users (containing 1k users) from two different directed and undirected social media application with blog discovery, post/comment discovery, and social network site connections.

We examine 5 different approaches for similarity relationships in social media applications. The approaches are

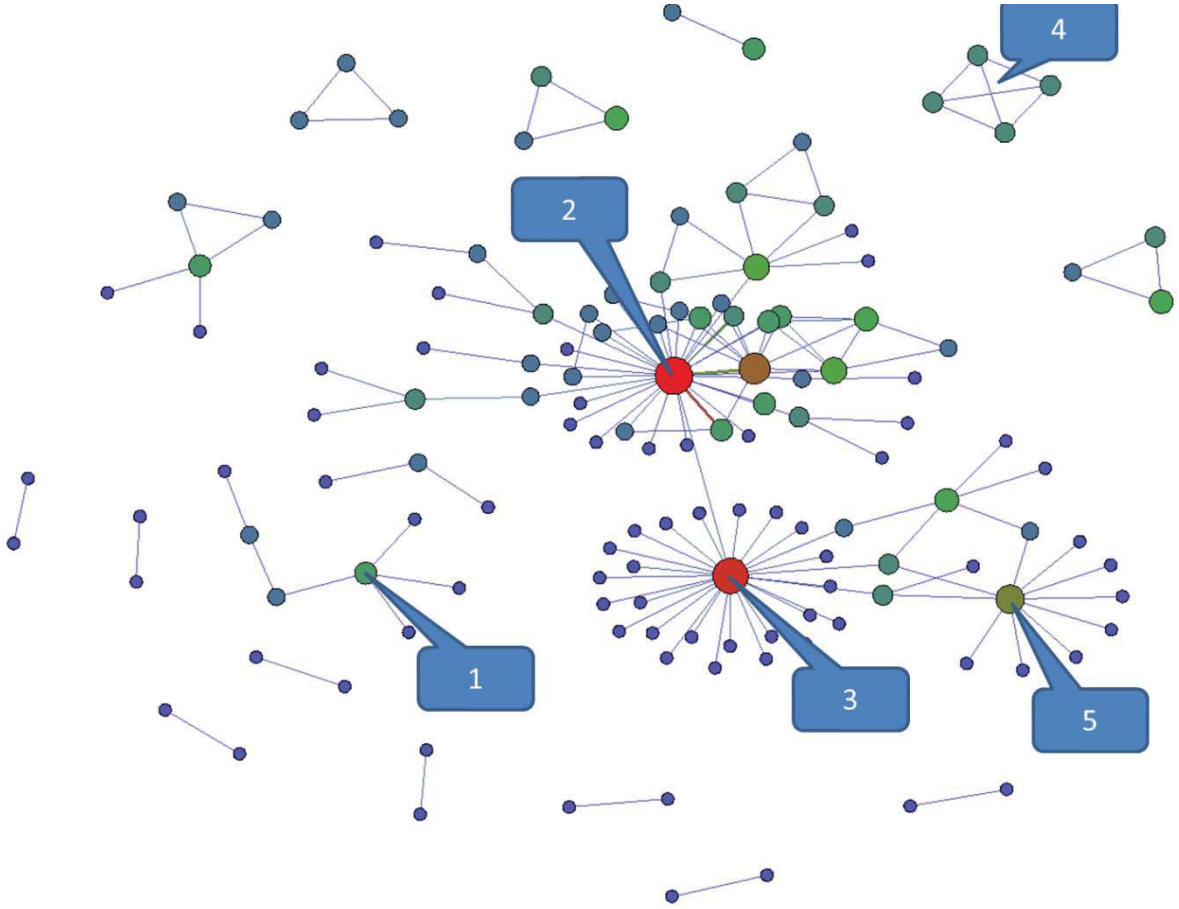
1.1 Active friend or network:

In this approach user similarity is accomplished based on the concept that two vertices are similar if their immediate neighbors in the network are themselves similar. This leads to a self-consistent matrix formulation of similar.



1.2 Spelling Distance:

The spelling distance metric as implemented by the SAS dataflux procedure is run on the user profile information for user similarity. This approach includes user matching on more than one set of fields, in which case records which match on either set of fields are claimed to be similar user to recognize strings that match inexactly but really represent the same person, address, etc., or two very different first names that are really a name and nickname. If user id's are standardized, most matching involves simple exact matches. Standardization will not correct typos, but will unify the different forms of a first name. Still, fuzzy matching remains very important. Without well-conceived fuzzy match algorithms, the result is likely to be unsatisfactory, with too many false positives and/or false negatives. Dataflux creates match codes for strings based on their characters, locale, and various user-selected settings. These settings include sensitivity – a measure of match code complexity.



1.3 TGPARSE Interval processing:

In this approach, incoming group of documents by users is scored using the text mining utilities of SAS Text Miner (specifically, Proc Docscore). Docscore is used to determine SVD values and append information to existing results. When the incoming document count reaches a determined number it is

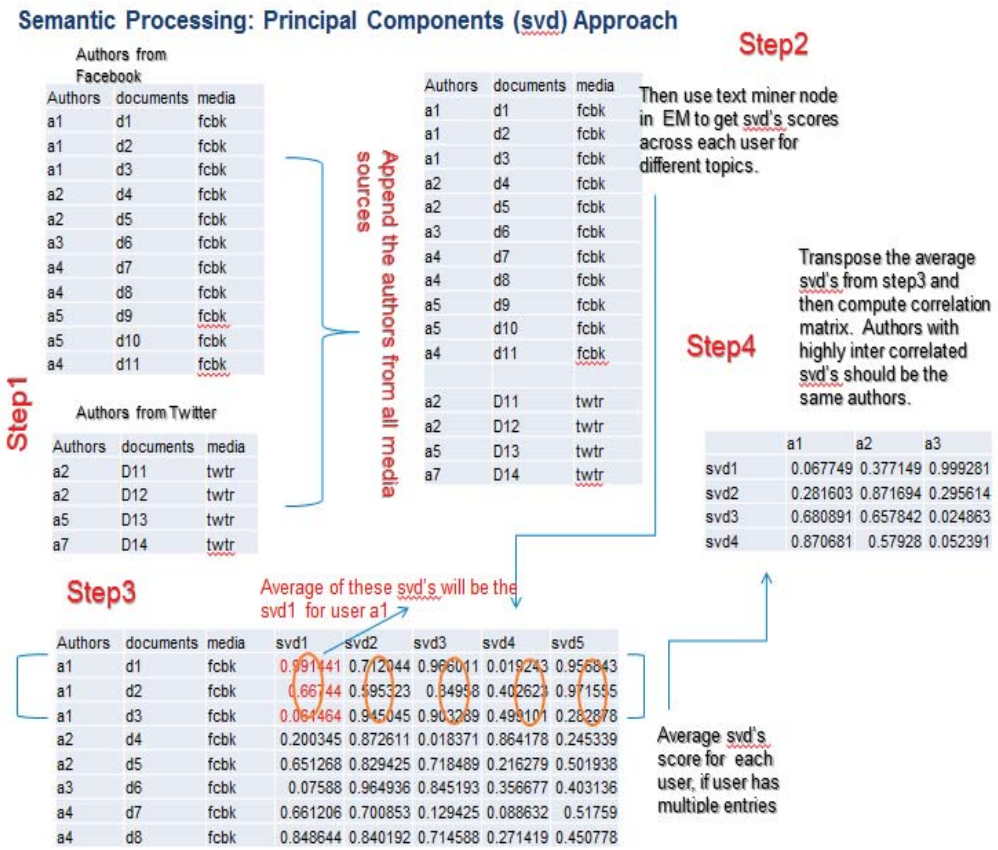
then assumed that the parse results have degraded and the Seed Method needs to be run again. This seed method would be run on a sample of documents – NOT on the entire document source. A Table will be maintained to monitor the degradation of the analytic methods used. This will trigger those methods to use the original seed method on a sample of the data and then the interval methods will continue

1.4 Principal Clustering:

In this approach we used clustering to group users based on principal components. Users within the same cluster are believed to be the same person. Incremental – new documents are scored individually using docscore and their principal component value is calculated. They are then placed into the existing cluster based on their value.

1.5 Eigenvalue clustering:

This approach is the same as #4 only Eigenvalues are used for clustering instead of principal component values. This approach assumes that a linear combination of terms in the documents are capable of explaining the variability of all the documents by users..'



2. Summarization of approaches

The number of similarity relationships that can be inferred from social media sources is diverse and depends on various factors, such as the level of adoption of the different applications within the organization, the frequency of activity that yields similarity (e.g., commenting on a blog entry vs. joining a community), and the likelihood of similar activities by other users (e.g., other users commenting on the same blog entry vs. using the same tag). As the potential for inferring similarity relationships is an important characteristic of the approach, and may affect its selection for certain scenarios, we inspect the number of relationships that can be inferred for each approach.

3. Harvesting and Aggregating Similarity Relationships

Rank is determined by a similarity score, which expresses the similarity strength between two individuals and is in the range of [0,1]. Similarity score is calculated for all sources using Jaccard's index, i.e., by dividing the number of items in the intersection set by the number of items in the union set. For example, similarity of bookmarks is the number of pages bookmarked by both users divided by the number of distinct pages bookmarked by any of these two users; for tagged with, the number of

tags both users are tagged with is divided by the number of distinct tags any of the users is tagged with.

Aggregations of different approaches by calculating a weighted average of their similarity scores to generate an aggregated similarity score in the [0,1] range. we examined each source separately, as well as aggregates of sources according to the people, things, and group categories, and an aggregate of all five approaches.

4. CHARACTERIZING SIMILARITY SOURCES

Throughout this proposal, we focused that mining similarity relationships from social media will have great value for a variety of scenarios. This hypothesis stems from the fact that mining familiarity relationships has shown great value. Our second hypothesis is that different similarity approaches hold different information and provide different value. Before we can examine the value of approaches in various scenarios, we must show that:

(1) Similarity relationships are uniquely different from familiarity relationships (to prove that we are creating new value and not reusing old value from familiarity)

(2) Certain types of similarity sources are uniquely different from other similarity approaches (to prove that similarity approach provide different results, and thus their aggregation may be useful for different tasks/scenarios)

5. Aggregation of results

We ran number of tests on each of all five methods method to determine reliability weights to use for each analytic method i.e on sample counts of true acceptance when null hypothesis says that users are similar. Example, if method1 proves to be much more reliable than method 5 it will have a higher weight when the results are compared.

Probability that each combination of users is similar=
(A1_weight*A1)+(A2_weight*A2)+….(A5_weight*A5)/ (A1_weight+A2_weight….+A5_weight).

Analytics Method	U1=U5	U5=U1	UN=UX
A1	1	0	1
A2	0	1	0
A3	1	0	0
A4	0	1	0
A5	1	0	0

6. CONCLUSION

In this research, we examine mining five different methods of similarity relationships in social media applications. Inspecting the data of 1k social media avid users, the similarity approaches produce lists of similar people that differ substantially from lists produced from familiarity sources. This is a notable finding, as familiarity sources are a current focus of research, and this suggests that similarity approaches provide unique data that may assist a variety of scenarios for which familiarity lists are not appropriate. Furthermore, similarity approaches produce very diverse lists, which suggest that aggregating them may produce richer results. Examining the similar characteristics among sources reveals that categorizing them according to people, things, and groups, is productive. An experiment featuring 1k avid users of social media extends the mining results. The user experiment evaluates similarity approaches for three different scenarios: blog discovery, post/comment discovery, and social network site connections. For each of the three scenarios, the experiment conclusively shows that similarity evidence generate positive user interest in these social media tasks. The experiment highlights a particularly rich similarity approaches that is most useful across all of the scenarios. The experiment also shows that aggregating similarity approaches yields analogous richness. This is particularly useful since people, places, and group categories captures different kind of variability's. Among the categorical aggregates examined, things is most effective, probability each pair of user is similar. This indicates that users value things like tags and bookmarks for the four social media scenarios.

Both the mining results and the user experiment show that user similarity in social media applications has great value. It is clear that users have much interest in similar people that share their tags, bookmarks, friends, blogs, and communities.

7. Acknowledgment:

None of this work would have been possible without the help of many others, primarily: Barry Deville, James Cox.

SESSION

KNOWLEDGE ENGINEERING AND INFORMATION FUSION + KNOWLEDGE DISCOVERY + SEMANTIC WEB + DATA MINING

Chair(s)

TBA

Stochastic and Chaotic Calculus Methods for Detection of Multi-Agile Threat Emitters

James A. Crowder, Raytheon
Intelligence, Information and Services, Aurora, CO, USA

Abstract— Current processing environments struggle with the detection and identification of wideband, multiply agile signals. They often rely on partitioning the data environment using Pulse Descriptor Word (PDW) attributes such as frequency, pulse width, and time. This process fragments wideband agile signals and downstream processing, such as residue processing, usually cannot recover all the pulses and form these signals. Agile signals may have their pulses scattered across multiple processors, making their recovery difficult. There may be many missing pulses which have been incorrectly claimed by other signals. This process is complicated because of dense, diverse signal environments. This work involves the use of higher order spectral techniques to detect pseudorandom patterns within the data stream (either single or dual stream), indicating widely agile signals are present in the environment. These techniques will be applied to the partitions resulting from the new front-end partitioning processes. The intent is to introduce the reader to stochastic and chaotic calculus methods and their applicability to complicated signal detection requirements and mission needs [3].

Keywords—Multi-Agile Emitters, ELINT, COMINT, Radar Detection, Knowledge Engineering, Information Fusion

1. INTRODUCTION: MULTI-AGILE SIGNAL DETECTION

Radar manufacturers utilize transmitters that are agile over a very wide frequency range in order to hide their signals among other signals and among noise in the environment. By widely varying the frequency pulse-to-pulse, it is hard to detect such signals, especially utilizing frequency coherency as most processing systems do. In addition, many transmitters use agilities in pulse width and pulse repetition frequency to further hide the signals. However, it is possible to use their agility characteristics against them. Figure 1 illustrates the frequency vs. time characteristics of one frequency agile signal. For real noise, there aren't man-made restrictions on spectral content or frequency excursion. For pseudorandom sequences utilized to drive widely agile frequency transmitters there are restrictions on the total bandwidth and total spectral content due to the pseudorandom sequences used to drive the transmitter. If we track the higher order Stochastic Derivatives of frequency vs. time characteristics of agile and noise signals, we find that they are very distinctive. .

We believe these techniques provide real promise for creating algorithms for use in the front-end of processing systems, to greatly enhance the capability to detect and characterize wideband multiply agile signals (see Figure 2). The application of the proposed Stochastic Calculus techniques provides a simple, but novel approach for partitioning data environments containing wideband agile signals. The higher order spectral techniques are a very innovative use of higher derivatives on spectral information that allows widely. We believe these techniques provide real promise for creating algorithms for use in the front-end of processing systems to greatly enhance the capability to detect and characterize wideband multiply agile signals. The application of the proposed Stochastic Calculus techniques provides a simple, but novel approach for partitioning data environments containing wideband agile signals. The higher order spectral techniques are a very innovative use of higher derivatives on spectral information that allows widely agile signals to be detected and characterized in dense, wideband environments with either single or dual stream data.

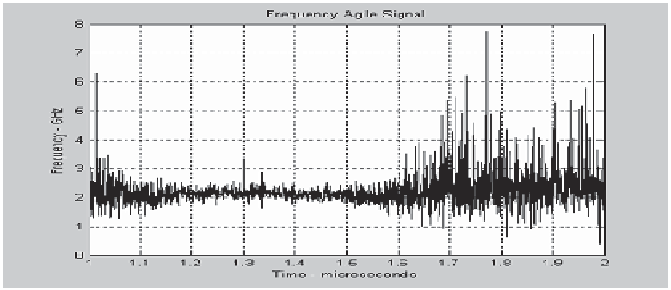


Figure 1 – Example of Freq. vs. Time for Wideband Agile Signal

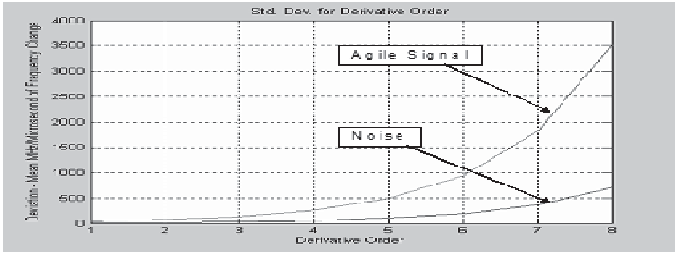


Figure 2 – Comparison of Higher Order Spectral Derivative for Agile vs. Noise Signals

2. THE STOCHASTIC DERIVATIVE

The fundamental properties of stochastic systems are important to the detection and characterization of wideband, agile emitters in a wideband data environment. Wideband, agile radars represent dynamical stochastic processes, since the signals are put through non-linear, stochastic processes imposed by the earth's environments [5]. However, the individual parametric changes within the radars (e.g.,

frequency and pulse-width), represent deterministic physical processes (i.e., they are driven by deterministic, pseudorandom generators). If not handled properly, the dynamical nature of the signals can cause a total divergence of system models and make these signals undetectable by standard techniques.

Stochastic Derivatives can be utilized to differentiate "pseudorandom" processes from actual random processes in a pulse environment. The problem is that Stochastic Derivatives are computationally challenging in a processing environment. What follows is a derivation of the Stochastic Derivative, ending in an approximation of the Stochastic Derivative utilizing an *L_2-space* Stochastic Derivative Matrix that greatly reduces the computational complexities of Stochastic Derivatives and produces and *O(n)* process, based on a discretized generalization of the Levy process.

2.1 The Stochastic Integral Reqpresentation

We start with a stochastic integral representation for an arbitrary random variable in a general L_2-continuous martingale space as an integrator for the process. Then, in relation to this, we will define a *Stochastic Derivative*. Through the derivative it can be seen whether the proposed random variable admits "or can be characterized" by the derived *Stochastic Integral* representation [8]. This allows us to show that the Stochastic Derivative determines the integrand for the Stochastic Integral which serves as the best *L_2-approximation* for the variable and allows us to establish a discrete *O(n)* approximation [1], called a ***Stochastic Derivative Matrix***.

We start by introducing a stochastic integration scheme in *L_2-space*:

$$H = L_2(\Omega, \Im, P)$$

for the class of real random variables ξ:

$$\|\xi\| = \left(E|\xi|^2\right)^{1/2}$$

and then introduce an *H-continuous* martingale process for integration, η_t, $0 \le t \le$ T, with respect to an arbitrary filtration of the process:

$$\Im_t, \quad 0 \le t \le \mathrm{T}$$

The integrands are considered as elements of a certain *functional L_2-space* of measurable stochastic functions:

$$\varphi = \varphi(\omega, t), \rightarrow (\omega, t) \in \Omega \times (0, T]$$

with a norm:

$$\|\varphi\|_{L_2} = \left(\iint_{\Omega\times(0,T]} |\varphi|^2 P(d\omega) \times d[\eta](\omega)\right)^{1/2} = \left(E\int_0^T |\varphi|^2 d[\eta]_t\right)^{1/2}$$

which is given by a "*product type*" measure:

$$P(d\omega) \times d[\eta]_t(\omega)$$

associated with a stochastic function $[\eta]_t, 0 \le t \le T$, having monotone, right-continuous stochastic trajectories, such that:

$$E(\Delta[\eta] \mid \Im_t) = E\left(|\Delta\eta|^2 \mid \Im_t\right)$$

for the increments $\Delta[\eta]$ and $\Delta\eta$ on intervals:

$$\Delta = (t, t + \Delta t] \subseteq (0, T]$$

In particular, for the *Levy process* η_t, $0 \le t \le$ T, as integrator:

$$\left(E\eta_t = 0, E\eta_t^2 = \sigma^2 t\right)$$

the deterministic function:

$$[\eta]_t = \sigma^2 t$$

is applicable. In particular, those functions having their permanent $\Im_t$ *-measurement* values $\varphi^h \in H$ on the *h-partition* intervals:

$$\Delta = (t, t + \Delta t] \subseteq (0, T] \qquad \sum \Delta = (0, T]$$
$$(\Delta t \le h)$$

have their stochastic integral defined as:

$$\int_0^T \varphi^h d\eta_s \overset{def}{=} \sum_{\Delta} \varphi^h \cdot \Delta\eta$$

Broken into the partition intervals associated with signal environments, it is assumed that

$$E\left(\varphi^h \Delta\eta\right)^2 = E\left(\left|\varphi^h\right|^3 \cdot E\left(\left|\Delta\eta\right|^2 \mid \Im_t\right)\right) =$$

$$E\left(\left|\varphi^h\right|^2 \cdot E(\eta) \mid \Im_t\right) = E\int_{\Delta}\left|\varphi^h\right|^2 d[\eta]_S < \infty$$

which yields:

$$E\left(\int_0^T \varphi^h d\eta_S\right)^2 = E\int_0^T \left|\varphi^h\right|^2 d[\eta]_S$$

where the *integrands* φ are *identified* as the limit:

$$\varphi = \lim_{h\to 0} \varphi^h$$

For the functional signal space, which we will define as an *L_2-space*: $\left\|\varphi - \varphi^h\right\|_{L_2} \to 0$, the corresponding *Stochastic Integrals* are defined as limits:

$$\int_0^T \varphi d\eta_S = \lim_{h\to 0} \int_0^T \varphi^h d\eta_S$$

in the measure space of *H*, with:

$$\left\|\int_0^T \varphi d\eta_S\right\| = \|\varphi\|_{L_2}$$

In our system of radar signals, we are assuming a "*Simple Stochastic Structure*" that represents the non-agile part of the signal environment. Given this, the integrands of the stochastic space can be characterized in our an *L_2-space* as functions φ on the non-Euclidean space $\Omega \times (0, T]$ which are measurable with respect to the σ-algebra generated by measurable spaces of the form: $A \times (t, t + \Delta t]$, with $A \in \Im_t$. If the measure space account for all signals within the environment, they will constitute a *semi-ring*, and the indicators of these *Stochastic Integral* equations will constitute a *complete system* within the *L_2-subspace* of functions, measurable with respect to the σ-algebra generated [2]. Since we are utilizing Levy processes as the integrator, this signal environment characterization can be simplified by

identification of the integrands as the *Stochastic Function:* φ, having $\Im_t$ *-measurable* values φ_t, $0 \le t \le T$:

$$\int_0^T \|\varphi\|^2 dt < \infty$$

From here we to see if the hypotheses set forth above hold, i.e., whether random variables established by the wideband radar environments:

$$\xi \in H$$

allows representation by the *Stochastic Integral Equation* established in Equation 3 and whether an integral approximation to ξ:

$$\hat{\xi} = \int_0^T \varphi d\eta_S$$

can be determine, where $\hat{\xi}$ is meant to be the projection of ξ onto the subspace $H(\eta)$ of all stochastic integrals with the considered Levy stochastic integrator η_t, $0 \le t \le T$. In particular, whether the integrand φ can be determined through the simple integrands φ^h described in above [4]. Here we will also consider the case within the subspace $H(\eta)$ of random variables $\hat{\xi} \in H$, where the *stochastic integral representation* is of the form [13]:

$$\hat{\xi} = \sum_{k=1}^{n} \int_0^T \varphi^k d\eta_s^k$$

with respect to the system of orthogonometic martingale processes:

$$\eta_t^k, \quad 0 \le t \le T \qquad (k = 1, \ldots, n)$$

as integrators.

2.2 Stocastic Derivatives and L_2 Approximations with Stochastic Integrals

Now we show that a Stochastic Derivative can be well defined for any "*Simple Stochastic*" process. Since the signal environment contains either Wideband Agile signals (which may or may not be multiply agile) and radars that are not Wideband Agile, Pulses that are in the environment that are not from a Wideband Agile signal are seen to form an *H-continuous* martingale process η_t, $0 \le t \le$ T. For the random variable $\xi \in H$, we define now its *Stochastic Derivative* $D\xi$ with respect to the integrator we have defined, η_t, $0 \le t \le$ T as [17]:

$$D\xi \overset{def}{=} \lim_{h \to 0} E\left(\xi \frac{\Delta\eta}{\|\Delta\eta\|_t^2} \mid \Im_t \right)$$

In particular, since we have defined our functions to be of the type specified in Equation 2, we more precisely define $D\xi$ as [15]:

$$D\xi \overset{def}{=} \lim_{h \to 0} \sum_{\Delta} E\left(\xi \frac{\Delta\eta}{\|\Delta\eta\|_t^2} \mid \Im_t \right) 1_{\Delta}(s) \quad 0 \le s \le \mathrm{T}$$

Giving us values for φ^h defined as:

$$\varphi^h = E\left(\xi \frac{\Delta\eta}{\|\Delta\eta\|_t^2} \mid \Im_t \right)$$

which is valid on the signal environmental partitions (*h-partition*) intervals $\Delta = (t, t + \Delta t]$ where :

$$\|\Delta\eta\|_t^2 = E\left(|\Delta\eta|^2 \mid \Im_t \right)$$

This allows us to show that the *Stochastic Derivative* (from Equations 5 and 6) is well defined and calculable for any $\xi \in H$, and has a unique *Stochastic Integral* representation as [6]:

$$\xi = \xi^0 + \int_0^T D\xi d\eta_s$$

Given that we have monotone *h-partitions* (as discussed above), for the subspace:

$$H(\eta) \subseteq H$$

For our *Stochastic Integral Equations* (shown in Equation 4) we have a limit (convergence) for the orthogonometric sums with their components $H(\Delta\eta)$ as [14][1]:

$$H(\eta) = \lim_{h \to 0} \oplus H(\Delta\eta)$$

Now we look at the stochastic variables in *H* of the form:

$$\psi \cdot \Delta\eta$$

where ψ are the $\Im_t$ *-measurable* multiplicators for the increments $\Delta\eta$ on the *h-partitions* $\Delta = (t, t + \Delta t]$.

Now we project ξ onto $H(\Delta\eta)$ with the multiplicator $\psi = \varphi^h$ with φ^h as defined in Equation 6. This provides us with the orthogonometric condition:

$$E\left(\xi - \varphi^h \Delta\eta\right)(\psi \Delta\eta) = 0$$

Therefore, the projections of ξ onto $H(\Delta\eta)$ are:

$$\sum_{\Delta} \varphi^h \Delta\eta = \int_0^T \varphi^h d\eta_s$$

Based on the equations above, this allows the projections $\hat{\xi}$ of ξ onto the subspace $H(\eta)$ for the orthogonometic stochastic differential pairs to be represented by a particular *Stochastic Limit Integral*:

$$\hat{\xi} = \int_0^T \varphi d\eta_s = \lim_{h \to 0} \int_0^T \varphi^h d\eta_s$$

[1] Orthogonometric refers to orthogonal pairs of stochastic and chaotic equations/processes.

since this is defined on H and we are using the *Stochastic Integrand* φ which, according to Equation 2, is the limit of φ^h, according to the equation:

$$\left\| \int_0^T \varphi d\eta_s - \int_0^T \varphi^h d\eta_s \right\| = \left\| \varphi - \varphi^h \right\|_{L_2}$$

which makes the representation in Equation 7, which is based on the *Stochastic Integrand* $\varphi = D\xi$, along with the difference equation:

$$\xi^0 = \xi - \int_0^T \varphi d\eta_s$$

orthogonal to the subspace $H(\eta)$.

This allows us to treat the Wideband Pulse environment as a set of *orthogonometric martingale processes*. The approximations of the *orthogonometric martingale processes* $\hat{\xi}$ s are computed as:

$$\hat{\xi} = \sum_{k=1}^{N} \int_0^T \varphi^k d\eta_s^k$$

with the *Stochastic Integrands*:

$$\varphi^h = \lim_{h \to 0} E\left(\xi \frac{\Delta \eta^k}{\left\| \Delta \eta^k \right\|_t^2} \mid \Im_t \right)$$

which now defines the *Stochastic Derivative* theory we need for use with the Wideband Pulse environments.

2.3 Stochastic Convergence and Fractional Derivatives

Now that we have established that the Wideband Pulse Environment can be treated as a set of *Orthogonometric Martingale Functions* (OMF), we must investigate the differences between the derivatives for stochastic vs. deterministic processes.

For this we look at derivatives with respect to a given σ-*field* (projection), $\hat{\xi}$, with our martingale processes posed as diffusion processes. By posing the problem in this framework, we can focus our efforts on the case where ξ is a fractional diffusion, or a partial representation of the infinite martingale processes (for our use this is a given collection of pulses over a finite period of time) [16]. In this case, $\hat{\xi}$ represents the past, the future, or the present of ξ. We portray this as a fractional derivative in the sense of the theory of distributions in order to examine the use of stochastic vs. deterministic convergence in order to differentiate between random and pseudorandom processes within the Wideband data environments [7].

The problem is to create an *O(n)* process which contains the dynamical meaning of an ordinary differential equation and which allows us to extend this dynamical meaning to stochastic and pseudostochastic processes to form orthogonometric pairs of fractional derivative pairs to fully capture the Wideband Data environment and determine the presence of Wideband, agile emitters within the pulse environment [8]. Unfortunately, the martingale processes represented by the dense, Wideband pulse environments which represent a different set of radars on any given time period, the limit:

$$\frac{\hat{\xi}_{t+k} - \hat{\xi}_t}{k}$$

does not really exist. The reconciliation of this limit to our problem lies in removing the divergences which would appear trying to correlate over many collection partitions and average over a series of orthogonometric martingale projections, studying the behavior when *k* goes to zero over the conditional expectation with respect to a *n-dimensional* Brownian fractional diffusion. For a given time t, we calculate a backward fractional derivative with respect to the σ-*field* (projection), $\hat{\xi}$, which is viewed as the past process up to the time *t*. Since we are posing our environments as fractional martingale processes, we look at theory pertaining to fractional Brownian motion, since this is the closest approximation of the Wideband pulse environments [18].

2.4 Fractional Brownian Motion

Here we define $\hat{\xi} = \left(\hat{\xi}_t\right)_{t \in [0,T]}$ to be a fractional Brownian motion process on $H \in (0,1)$ and is defined on a stochastic space (Ω, F, P) with a covariance function:

$$E\left(\hat{\xi}_s \hat{\xi}_t\right) = R_H(s,t)$$

where:

$$R_H(s,t) = \int_0^{s \wedge t} \varphi^H(s,u) \varphi^H(t,u) du$$

where φ^H defined by:

$$\varphi^H(s,t) = c_H s^{1/2-H} \int_s^t (u-s)^{H-3/2} u^{H-1/2} du \quad 0 < s < t$$

and where:

$$c_H^2 = H(2H-1)\beta(2-2H, H-1/2)^{-1}$$

and β denotes a Beta function we set $\varphi^H(s,t) = 0, if\ (s \geq t)$. Now we let:

$$\varphi^{H^*} : \varepsilon \to L^2([0,T])$$

be the linear operator defined by:

$$\varphi^{H^*}\left(1_{[0,t]}\right) = \varphi^H(t,\cdot)$$

Then the following equality would hold for any $\phi, \psi \in \varepsilon$:

$$\langle \phi, \psi \rangle_H = \left\langle \varphi^{H^*}\phi, \varphi^{H^*}\psi \right\rangle_{L^2([0,T])} = E\left(\hat{\xi}(\phi)\hat{\xi}(\psi)\right)$$

Allowing φ^{H^*} to provide an isometry between the Hilbert Space *H* and a closed subspace of $L^2([0,T])$ and allows us to create our orthognometric space for pseudorandom/random martingale differential pairs within the signal pulse environments [9].

2.5 Space of Deterministic Stochastic Integrands

For our signal environment subspaces, the following assumptions are made [10]:

- The number of non-agile radars is bounded: $< \infty$
- The number of agile radars is bounded: $< \infty$
- *H*, the space of non-agile radars, is > ½ (less than ½ of all the pulses captured are from agile radars).
- Agile radars are driven by deterministic processes (either pseudorandom or deterministically chaotic) and therefore their derivatives in stochastic subspace are bounded
- Non-agile radars form a stochastic environment of martingale processes that are bounded and therefore their stochastic derivatives are bounded.

Given these assumptions, we look at the set of *Stochastic Differential Equations* driven by $\hat{\xi}$. If we look at the set of random variables (*S*) derived from the environment, for any given agile emitter, its pulses would have gone through the same random transformation corresponding to the physical environment (e.g., atmospheric conditions, etc.). This differentiates from the random variable posed by the other radars within the signal environment, forming a system of random variables, modeled as an orthogonometic set of martingale processes $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_n$, where *n* is bounded [11]. If we let ϕ denote the environments, then we can express the set of functions *F* as:

$$F = f\left(\hat{\xi}(\phi_1), \ldots \hat{\xi}(\phi_n)\right)$$

for $n \geq 1$, where the convergence is "*smooth*" and where $f : \Re^n \to \Re$ is a smooth, compact function and each function belongs to the Hilbert space *H*, or $\phi_i \in H$. Therefore, the *Stochastic Derivative* of *F*, with respect to $\hat{\xi}$ is the element of $L^2(\Omega, H)$ defined by:

$$D_s^{\hat{\xi}} F = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\left(\hat{\xi}(\phi_1), \ldots, \hat{\xi}(\phi_n)\right)\phi_i(s) \text{ for } s \in [0,T]$$

In particular, $D_s^{\hat{\xi}} \hat{\xi}_t = 1_{[0,t]} s$, and $D^{1,2}$ denotes the closure of this set of random variables (convergence) with respect to the L^2 norm we have defined:

$$\|F\|_{1,2}^2 = E\left[F^2\right] + E\left[\left|D^{\hat{\xi}}\right|_H^2\right]$$

Therefore the *Stochastic Derivative* $D^{\hat{\xi}}$ has a stable chain rule and $(F_i)_{i=1,\ldots n}$ is a sequence of elements of $D^{1,2}$ such that for any $s \in [0,T]$,

$$D_s^{\hat{\xi}} \varphi(F_1, \ldots, F_n) = \sum_{i=1}^{n} \frac{\partial \varphi}{\partial x_i}(F_1, \ldots, F_n) D_s^{\hat{\xi}} F_i$$

We can then define a "*divergence operator*" $\delta^{\hat{\xi}}$, which is the adjoin of the derivative operatory $D^{\hat{\xi}}$ which can be used to determine whether the backward *Stochastic Derivative* is heading toward *Stochastic Convergence* or *Stochastic Divergence*. In particular, if a random variable $\hat{\xi} \in L^2(\Omega, H)$ belongs to the domain of the divergence operator $\delta^{\hat{\xi}}$, it will satisfy:

$$\left|E\left\langle D^{\hat{\xi}} F, i\right\rangle_H\right| \leq \hat{\xi}_i \|F\|_{L^2} \quad \text{for any } F \in S$$

Therefore, for any of our martingale sequences, $\hat{\xi}$, that are moving toward *Stochastic Convergence* as the discrete, real-time, backward *Stochastic Derivatives* are being computed, $\delta^{\hat{\xi}}(i)$ is defined by the duality relationship:

$$E\left(F\delta^{\hat{\xi}}(i)\right) = E\left\langle D^{\hat{\xi}} F, i\right\rangle_H \quad \text{for every } F \in D^{1,2}$$

or, for those functions that exhibit *Stochastic Convergence*, each of their Stochastic Derivative orders will also exhibit *Stochastic Convergence* [12]. Therefore, we only need investigate, compute, and correlate the fractional, backward, *Stochastic Derivatives* of the Signal Environment to determine whether agile signals are present in a given pulse environment collection.

3. RESULTS

In order to properly test the Stochastic Derivative algorithms and MATLB Code, Wideband pulse environments have to be created. The assumptions for these environments are:

- Most of the pulse data environments, > 95%, include non-agile radars across the wideband frequency range.
- There are few (3 or less) wideband, agile radars in the pulse environments.
- Each agile radar has less than 200 pulses within the pulse environment to be tested

3.1 Generating Agile Radar Pulse Parameter Sequences

Three multiply agile radar signals were created for the initial testing of the Stochastic Derivative algorithms. Each was widely agile in frequency and pulse-width. Figures 3-5 illustrates the frequency PMFs for the three agile emitters. Figures 6-8 illustrate the pulse-width PMF profiles for the three agile emitters.

3.2 Generative Non-Agile Radar Pulse Parameter Sequences

In order to test the effectiveness of detecting Wideband, agile radars in dense pulse environments, it was necessary to create a series on non-agile radars and then interleave their pulses with the Wideband, agile radar pulses in order to create PDW environments that would represent data environments like are found in processing systems. In order to facilitate this kind of testing, 100 non-agile radars were generated (their parametric sequences were generated). Figure 9 shows the frequencies for each of the 100 non-agile radar pulses. Figure 10 provides a plot of the pulse-widths for the non-agile radars.

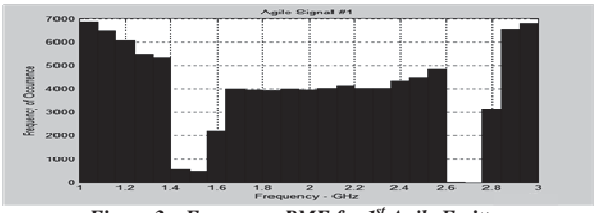


Figure 3 – Frequency PMF for 1st Agile Emitter

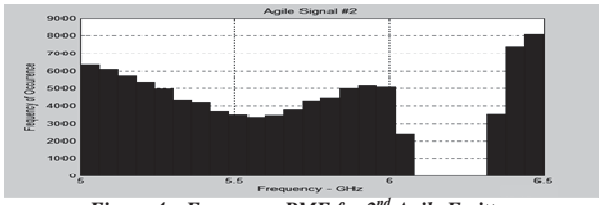


Figure 4 – Frequency PMF for 2nd Agile Emitter

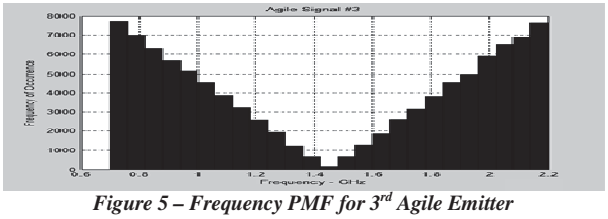


Figure 5 – Frequency PMF for 3rd Agile Emitter

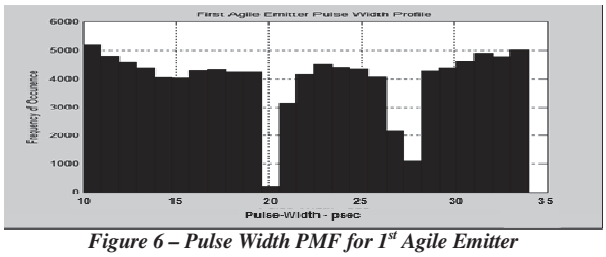


Figure 6 – Pulse Width PMF for 1st Agile Emitter

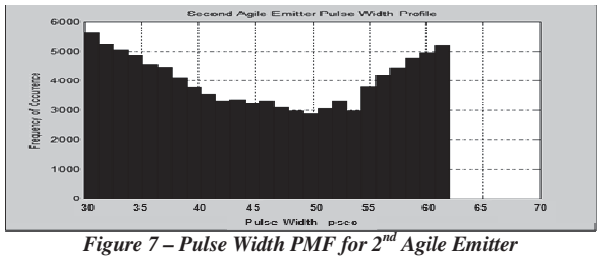


Figure 7 – Pulse Width PMF for 2nd Agile Emitter

3.3 Radar Environments

In order to test the Stochastic Derivative algorithms, radar pulse environments were created, utilizing the algorithms described above. Pulse environments were created with the following percentages of wideband, multiply-agile signals in the signal environment:

- 0%
- 0.15% with one wideband, multiply-agile signal
- 0.20% with 20 pulses out of 10,000 from two wideband, multiply-agile signals
- 0.50% with 50 pulses out of 10,000 from two wideband, multiply-agile signals
- 3.00% with 300 pulses out of 10,000 from three wideband, multiply-agile signals

Figures 11 and 12 illustrate examples of the 0.15% radar parametric environments to be tested. Figure 11 illustrates the frequencies for the first 100 pulses and Figure 12 shows the pulse widths for the first 100 pulses

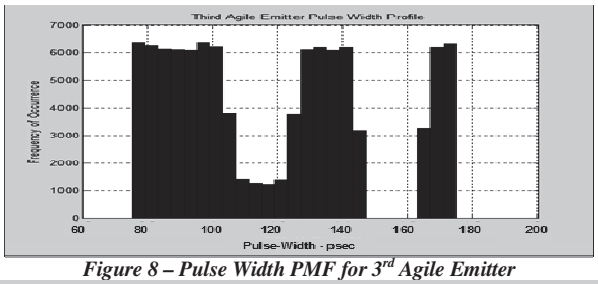


Figure 8 – Pulse Width PMF for 3rd Agile Emitter

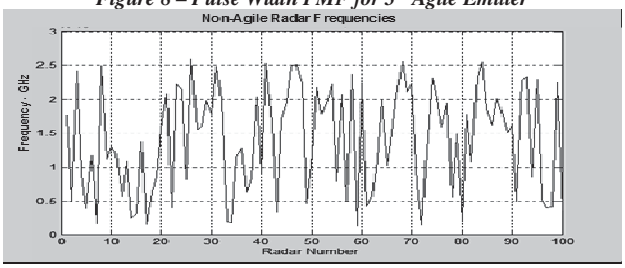


Figure 9 – Non-Agile Radar Frequencies

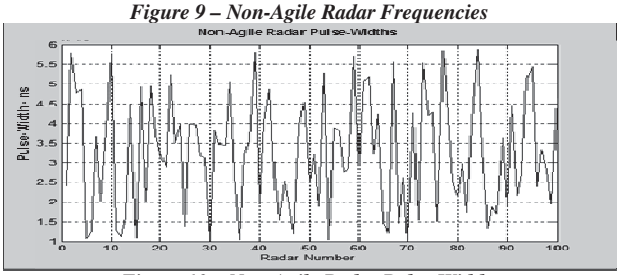


Figure 10 – Non-Agile Radar Pulse-Widths

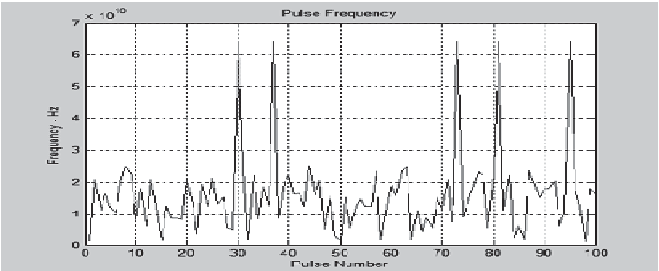


Figure 11 - Frequencies for the 1st 100 pulses of the 0.15% agile signal environment

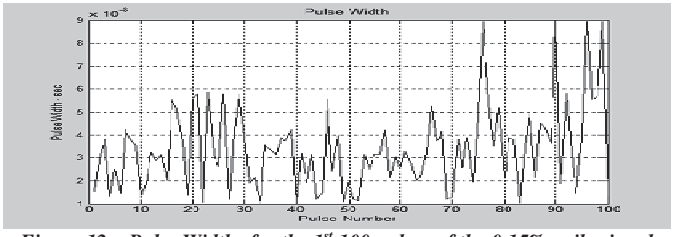


Figure 12 – Pulse Widths for the 1st 100 pulses of the 0.15% agile signal environment

3.4 Non-Agile Radar Pulse Environments

Figures 13 and 14 illustrate the frequencies and pulse widths for the radar signal pulse environment created with no wideband, multiply-agile emitters

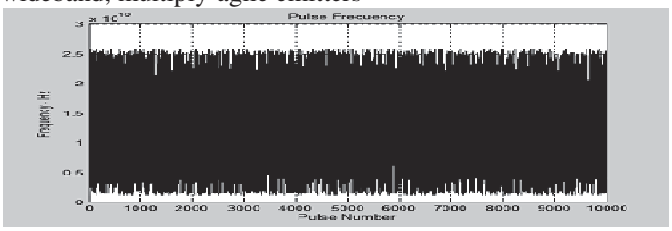


Figure 13 - Frequency Plot for Non-Agile Radar Signal Pulse Environment

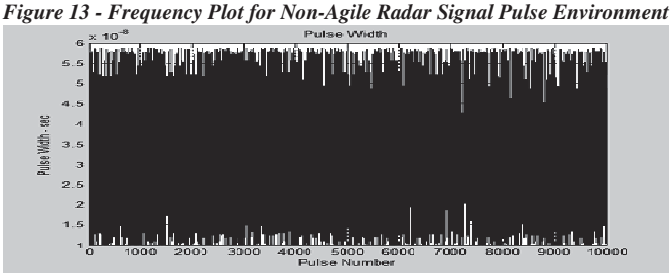


Figure 14 - Pulse-Width Plot for Non-Agile Radar Signal Pulse Environment

3.5 Testing Results

Figure 15 illustrates the results of utilizing the Stochastic Derivative algorithms on the pulse environments described above. The Higher-Order moments of the 1st 8 Stochastic Derivatives were generated and plotted for each of the signal environments. As can be seen from 15, even 15 pulses from a pseudorandom-driven, multiply-agile radar signal caused a jump in the Stochastic Derivative moments. And 300 pulses out of 10,000 cased a major jump in the Higher Order Stochastic Derivative moments. Clearly the algorithms can detect the presence on non-stochastic signals in the environment.

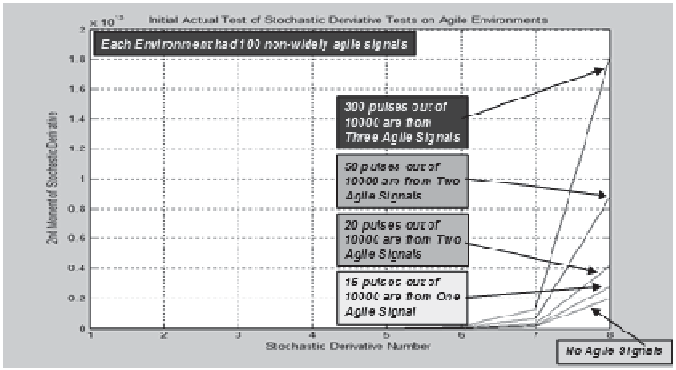


Figure 15 - Results of Stochastic Derivative Tests on the Signal Environments

4. REFERENCES

1. A.F. Faruqi, K.J. Turner, Appl. Math. Comput. 115 (2000) 213.
2. Baudoin, F. and Nualart, D. (2003). Equivalence of Volterra processes. Stochastic Process. Appl. 107 327–350.
3. Biane, P., 1999. Chaotic Representations for Finite Markov Chains. Stochastics and Stochastic Reports, 39: 61-68.
4. Cresson, J. and Darses, S. (2006). Plongement stochastique des syst`emes lagrangiens. C. R. Acad. Sci. Paris Ser. I 342 333–336.
5. Crowder, J. A. 2002. Derivation of Fuzzy Classification Rules from Multidimensional Data. NSA Technical Paper CON_0013_2002_001.
6. Crowder, J. A., 2007 "Integrating Metrics with Qualitative Temporal Reasoning for Constraint-Based Expert Systems." Proceedings of the 2007 Processing Systems Technology Network Conference, El Segundo, CA.
7. Crowder, J. A., 2005 "Multi-Sensor Fusion Utilizing Dynamic Entropy and Fuzzy Systems." Proceedings of the 2004 PSTN Processing Technology Conference, Tucson, Arizona.
8. Crowder, J., Barth, T., and Rouch, R. 1999. Learning Algorithms for Stochastically Driven Fuzzy, Genetic Neural Networks. NSA Technical Paper, ENIGMA_1999_002.
9. Di Nunno G. (1999). On stochastic differentiation. Quaderno IAMI 99.23, CNR-IAMI, Milano.
10. Di Nunno G. and Rozanov Yu.A. (1999). On stochastic integration and differentiation. Acta Applicandae Mathematicae, 58, 231-235.
11. Doss, H. (1977). Liens entre ´equations diff´erentielles stochastiques et ordinaires. Ann. Inst. H. Poincar´e Probab. Statist. 13 99–125.
12. Emery, M., 2000. A Discrete Approach to the Chaotic Representation Property. 005, IRMA
13. F¨ollmer, H. (1986). Time reversal on Wiener space. Stochastic Processes—Mathematics and Physics (Bielefeld, 1984) 119–129. Lecture Notes in Math. 1158. Springer, Berlin.
14. Feinsilver, P., 1996. Some classes of Orthogonal Polynomials Associated with Martingales. Proceedings of the American Mathematical Association, 98(2): 298-302.
15. Leonov V. and Shiryaev A. (1959). On computation techniques of semiinvariants. Theory of Probability and its Applications, 3, 342-355.
16. Malliavin P. (1997). Stochastic Analysis. Springer-Verlag, New York.
17. Rozanov Yu.A. (2000) On differentiation of stochastic integrals. Quaderno IAMI 00.18, CNR-IAMI, Milano.
18. Schoutens, W., and Teugels, J., 1998. Lévy Processes, Polynomials, and Martingales. Commun. Statist. And Stochastic Models, 14:335-349.

A Semantic Knowledge Discovery Strategy for Translational Research

Cartik R Kothari

Department of Biomedical Informatics, School of Medicine, The Ohio State University, Columbus, OH

ABSTRACT

The identification of teams of researchers with widely differing areas of expertise who can collaborate as part of inter-disciplinary research teams is crucial to the success of translational research initiatives. We present a novel, semantic, metadata based knowledge discovery strategy, which identifies researchers from diverse research disciplines who can collaborate on translational research projects. The search strategy is novel in the sense it connects researchers based on their differences rather than on their similarities. This is different from conventional knowledge discovery strategies, which rely on similarities between data points to bring them together. Quantitative metrics associated with the results of this methodology indicate the viability of the inter-disciplinary research teams that were identified.

Keywords:

Metadata, Knowledge Discovery, Profiles, Translational Research, Semantics

1. Introduction

Recent scientific research initiatives such as the Human Genome Project have led to a data deluge. Hospitals, financial institutions, and social networks are all confronted with ever increasing volumes of data. This has led to at least two important insights. First, integrating the data from different domains of research and deriving meaningful information from it are crucial to solving long-standing research problems: from curing cancer to understanding the origins of life. Second, solving these research problems will require the collaboration of experts from many disciplines such as biology, computer science, statistics, finance, and law.

Centers for translational research aim to bring together scientists from different disciplines in an effort to identify novel approaches to address pressing research problems. Translational research by definition [1, 2] has had a clinical focus i.e., on improving patient healthcare by correlating human genomic (bench) data with clinical (bedside) data. We extend this definition to include collaborative opportunities focusing on other areas as well such as space exploration and replenishable (non-fossil based) energy sources.

If used to identify inter-disciplinary teams of researchers, conventional knowledge discovery techniques such as clustering would attempt to cluster researchers based upon their similarities. However, these techniques would not be very effective in bringing together people from different backgrounds. Instead, we have developed a non-statistical metadata based methodology that can connect researchers with widely differing backgrounds. In the following sections, we present the semantic metadata based knowledge discovery methodology used to identify potential collaborators for translational research and the results of the implementation of that methodology.

2. Materials and Methods

2.1 Background: Discovery Themes and the Data Analytics Collaborative

In 2013, The Ohio State University launched the Discovery Themes initiative (**discovery.osu.edu/about/guiding-principles.html**) as a transformative effort to identify novel research themes to address critical societal needs. The Discovery Themes initiative has an emphasis on collaborative research bringing together scientists from various disciplines and institutions. Data Analytics (**discovery.osu.edu/focus-areas/data-analytics/**), one of the identified Focus Areas of the initiative, will be responsible for "sifting through, organizing, and analyzing vast amounts of information and drawing conclusions based on that analysis." These conclusions will have an impact in diverse research areas: from smart materials to optimized healthcare costs. The Data Analytics Collaborative (DAC) at The Ohio State University has identified four areas of strength or "clusters" where Data Analytics can make a significant impact: a) Health and Well-Being, b) Climate and Environment, c) Complex Systems and Network Science, and d) Foundational Sciences [3].

The identification of collaboratory teams of researchers from diverse areas of expertise is central to the mission of the DAC. With this stated purpose, a research team was contracted to identify possible collaborations among researchers at The Ohio State University and other institutions.

2.2 Raw Data

Raw data for identifying cross-disciplinary research teams was gathered by compiling very large lists of research documents from the last 5 years, which included: a) funded research grant proposals from the web sites of the National Institutes of Health (NIH), the National Science Foundation (NSF), and the Office of Research at The Ohio State University, b) transcripts of TED talks from the TED (**www.ted.com**) web site, and c) abstracts of every article in every journal indexed by SCOPUS (**www.scopus.com**), restricting only to the top 10% of the most productive researchers in each discipline. The productivity of the researchers was estimated using the Faculty Scholarly Productivity Index (FSPI), a metric developed by Academic Analytics Inc (**www.academicanalytics.com**). Once this corpus was complete, a text mining procedure was used to compile a set of approximately 60,000 keywords, excluding stopwords (viz. connectives, pronouns, and articles). A researcher – keyword matrix was compiled relating approximately 15,000 researchers with specific collections of associated keywords.

Address for correspondence: Cartik R Kothari Saravanamuthu, Center for Biomedical Informatics, Harvard Medical School, Longwood Campus, Countway Medical Library, 10 Shattuck Street, Boston MA 02115. USA

Email: cartik_saravanamuthu@hms.harvard.edu

Each keyword was associated with 60 researchers on average, with some being associated with as many as 2500 researchers.

2.3 Preliminary Work

In a preliminary attempt to identify groups of interdisciplinary researchers, simple hierarchical clustering techniques [4] were implemented on the researcher – keyword matrix. However, the clustering was based only on common keyword usage among the researchers. All the keywords were weighted equally. For example, the keyword *"BRCA1,"* the name of a gene commonly implicated in cancer, which was associated with just 30 researchers, was weighted the same as *"Patient"* or *"Because,"* which were associated with thousands of researchers. Therefore, the quality of the "semantics agnostic" clusters obtained in this fashion was questionable. Moreover, clustering is an unsupervised machine learning technique that groups data points together based upon their similarity. Given the need to identify collaborating researchers from widely differing backgrounds, clustering was not the best technique to use. As a prerequisite to implementing more intelligent knowledge discovery algorithms, the keywords extracted in the raw data collections phase were preprocessed to add semantic content to them.

2.4 Data Preprocessing

The need to add semantics to the keywords required their categorization under different research disciplines or subheadings. Prior to this, the keywords underwent spelling corrections, splitting, and morphological lemmatization (§ Keyword Processing).

2.4.1 Keyword Processing

Starting with approximately 60,000 keywords, the following steps were followed.

1) Keywords with little information content (e.g. *"Because"* and *"Above"*) were eliminated. Simple text extraction methods fail to eliminate these words along with stop words such as connectives, prepositions, and articles. About 2000 words such as these were identified manually and eliminated from the collection.
2) Verbs, proper nouns, adjectives, adverbs, and all words other than common nouns were eliminated. The Stanford Natural Language Processing API (StanfordNLP) [5] was used to identify the parts of speech of every keyword.
3) Simple spelling mistakes were corrected using Norvig's spell checker [6]. The researcher lists associated with the correctly spelled keyword and the incorrectly spelled keyword(s) were condensed into one list corresponding to the correctly spelled keyword.
4) Keywords occur in various morphological forms. *"Oxidized," "oxidizes,"* and *"oxidizing"* are all morphological variants of *"oxidize."* Using the StanfordNLP, all such variants were mapped back (lemmatized) to their respective morphological roots. The researcher lists associated with each variant were also condensed into one list for the morphological root.
5) Some keywords were agglomerations of individual words such as "JoinedAtTheHip." A simple greedy search method split such keywords into their constituents and the constituents were lemmatized as well (step 4 above).
6) Since the knowledge discovery technique (described in the next sections) leveraged keywords that connected researchers together, keywords that were associated with only one author were also eliminated.

These steps condensed the starting keyword list to approximately 1600 keywords. The necessary semantics and semantic metrics were now added to each of these keywords.

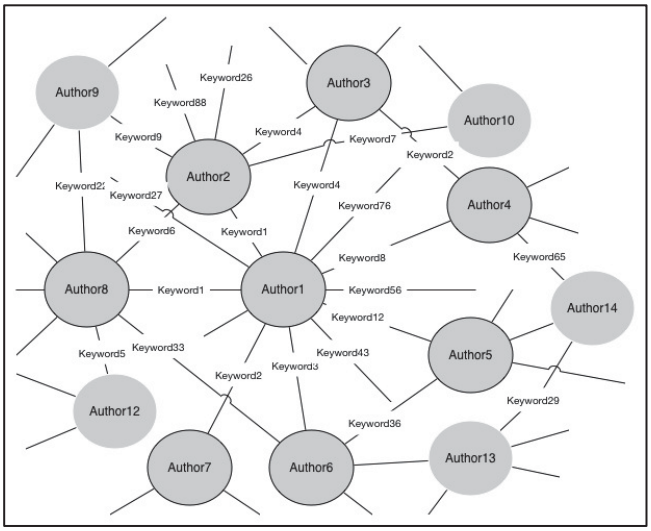


Figure 1:The Researcher – Keyword Hyperconnected Graph

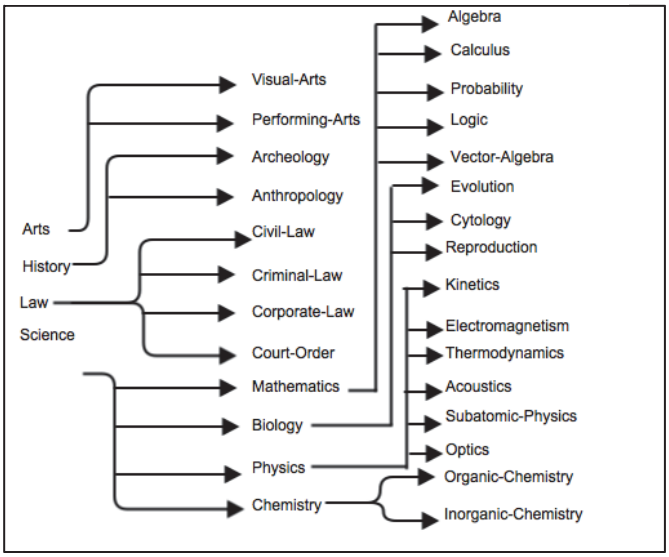


Figure 2 – An excerpt from the Subject Headings Hierarchy for Keyword Annotation

Given the researcher – keyword matrix now corresponded to a hyperconnected graph (Figure 1), the problem of identifying collaborations between the researchers is transformed into a graph search strategy. The addition of the semantic metrics significantly reduced the complexity of the search algorithm (§ The Knowledge Discovery Algorithm: The Path Finder).

2.5 Subject Headings Hierarchy for Keyword Annotation

The keywords were classified under various subject headings. Because the keywords were extracted from documents from different disciplines, we created a subject headings hierarchy loosely based upon a classification scheme used by the United States Library of Congress (www.loc.gov/aba/cataloging/classification/lcco/). We created an ontology [7, 8] of these subject headings. Figure 2 shows an excerpt from the hierarchies associated with these subject

headings. This classification scheme comprises 17 subject headings that are the roots of shallow 3-tier hierarchies. From the standpoint of semantic metrics (discussed in the next section), these hierarchies were deliberately kept shallow. Each keyword was classified under one of 17 subject headings viz. *Sport, Religion, History, Arts, Language/Literature, Law, Political Science, Education, Philosophy, Military, Psychology, Geography, Agriculture, Technology, Science, Medicine,* and *Social-Science*. Of these, *Medicine* and *Science* had the most number of subheadings. Most of the keywords were classified under these subject headings. Conversely, subject headings such as *Sport, Philosophy,* and *Religion* had the lowest number of keywords and no subheadings.

2.6 Semantic Metrics for Keywords

We used two different kinds of semantic metrics for the keywords. One was intrinsic to the keyword while the other was a measure of its distance from every other keyword. The intrinsic semantic metric for each keyword had to take into account: i) its specificity and ii) frequency of usage.

2.6.1 Specificity

Given the subheading ***SH*** under which a keyword was classified, specificity ***Sp*** is the length of the path from the subheading ***SH*** to the root of the subject heading hierarchy ***H SH*** occurs in. This is shown in (1). For example, the keyword *"Rice"* was classified under the *Crops* subheading (Figure 3), which is a part of the *Agriculture* subject heading, and its path to the root of this subject heading is given by *Crops:Plant-Culture:Agriculture*. The specificity measure of this keyword is given by 3, since its subheading *Vegetables* occurs at a depth 3 in the *Agriculture* subject heading hierarchy.

$$Sp_{SH}(Keyword) = Len(Path(SH, Root)); SH, Root \in H \quad (1)$$

2.6.2 Frequency of Usage

The usage count of every keyword was part of the input to the algorithm. Given the widely varying usage counts (from as low as 2 to several thousands of researchers), we adopted a logarithmic measure of usage frequency ***U***, with the median of the usage counts ***M*** being the base of the logarithm. The frequency measure ***Fr*** of a keyword is shown in (2).

$$Fr(Keyword) = Log_M(U) \quad (2)$$

2.6.3 Semantic Metric of the Keyword

The semantic metric of the keyword increases with increasing specificity but decreases with increasing usage counts. Therefore the semantic metric *SM* of the keyword is computed as the ratio of its specificity to its frequency (3).

$$SM(Keyword) = \frac{Sp(Keyword)}{Fr(Keyword)} \quad (3)$$

2.6.4 Semantic Distance between Keywords

The semantic distance *SD* between two keywords K_1 and K_2 is the sum of the depths of the subheadings SH_1 and SH_2 the keywords were classified under and the semantic distance between the subject heading hierarchies H_1 and H_2 they belong to. If the keywords belonged to the same hierarchy, the semantic distance between them was defined to be zero. This is shown in (4) and depicted in Figure 3.

$$SD_{K1,K2} = Sp(K_1) + Sp(K_2) + SD_{H1,H2}(K_1, K_2); H_1 \neq H_2$$
$$SD(K1, K2) = 0; H1 = H2 \quad (4)$$

The semantic distance between any two subject heading hierarchies was predefined based upon how closely related they are, i.e. how concepts in one heading overlap with another. Subject headings such as *History* and *Geography*

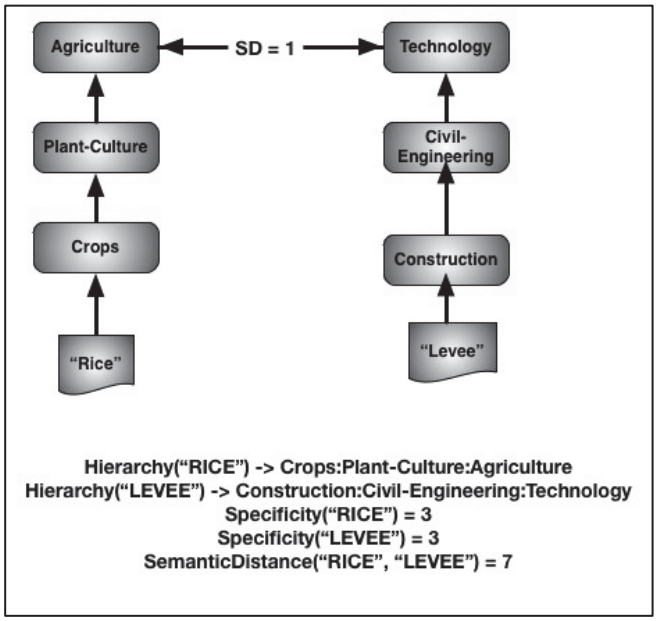


Figure 3:Specificity of the keyword "Rice"and its semantic distance from "Levee

are more closely related than *Art* and *Technology* for example; they have more common concepts. Therefore the semantic distance between *Art* and *Technology* was defined to be larger than between *History* and *Geography*. In this manner, we defined the semantic distances between every pair of subject headings to be one of two numeric values depending upon how closely related they are.

2.7 Knowledge Discovery Algorithm: The Path Finder

We implemented a graph search algorithm, a path finder, which generated paths across the researcher – keyword matrix. The researcher - keyword matrix can be visualized as a hyperconnected graph (Figure 1) with the researchers as the nodes and the keywords as the edges connecting them. The path finder algorithm was first implemented as a simple depth-first search [9] generating paths that included between 3 and 5 edges, i.e. the algorithm connected nodes (researchers) that were separated by between 3 and 5 keywords. Given the high cardinality (number) of edges from every node, a brute force path finder algorithm was very time consuming and generated millions of possible paths. The semantic metrics calculated for each keyword and the semantic distances computed for each keyword pair were used to reduce the search complexity. In addition, a third metric, the Path Score, was postulated to rank and filter the generated paths.

2.7.1 Path Score

The path scores *PS* were calculated for each path by summing the individual semantic metrics *SM* for each keyword *K* and also the semantic distance *SD* between every successive pair of keywords as shown in (5)

$$PS_{path} = \sum_{i=1}^{5} SM(K_i) + \sum_{i=1}^{4} SD(K_i, K_{i+1}) \quad (5)$$

Three different thresholds were used in conjunction with these metrics to prune the search space of the path finder algorithm and also to rank the results.

1. First, a threshold, the Keyword Metric Threshold (KMT), was set such that only keywords with a semantic metric higher than this threshold were used in the path finder algorithm. The lower the set value of the KMT, the more keywords were explored, which increased the time and space resources of the path finding algorithm.
2. Next, a second threshold, the Semantic Distance Threshold (SDT), was set such that every keyword in the explored path had non-zero semantic distance from all the other keywords. In other words, each keyword in the 5-keyword path was constrained to come from a different subject heading (ref. (4)). This has the effect of making the search diverge more quickly, bringing together distantly related nodes faster.
3. Lastly, a Path Weight Threshold (PWT), was set such that only those generated paths whose path scores (ref. (5)) were higher than the PWT were considered.

Algorithm 1: Path Finder Algorithm

```
Input: Matrix(Author, Keyword) as akm
Input: Matrix(Keyword, Author) as kam
Input: DegOfSep, KwMetThresh, SemDistThresh, PathWtThresh
1.  Function PathFinder (akm, kam)
2.      Foreach (author) in akm
3.          Path <- new Path
4.          Path <- Path.append(author)
5.          extendPath(Path, akm, kam)
6.      Endfor
7.      Return
8.  Endfunction
9.  Function extendPath (Path, akm, kam)
10.     if (Path.getNumberOfKeywords == DegreesOfSep &&
11.         Path.getWeight > PathWtThresh)
12.         print(Path, Path.getWeight)
13.         Return
14.     Endif
15.     keywordColl <- getKeywords(author, akm)
16.     Foreach (keyword) in keywordColl
17.         if(keyword.getMetric > KwMetThresh &&
18.             getSemDist(keyword, prevKeyword) > SemDistThresh)
19.             authorColl <- getAuthors(keyword, kam)
20.             Foreach (author) in authorColl
21.                 Path <- Path.append(keyword)
22.                 Path <- Path.append(author)
23.                 extendPath(Path, akm, kam)
24.             Endfor
25.         Endif
26.     Endfor
27. Endfunction
```

These thresholds significantly reduced the complexity of the path finder algorithm in terms of time and the number of valid generated paths. The steps of the Path Finder algorithm are outlined in Algorithm 1 above.

3. Results

We implemented the path finder algorithm with the thresholds and the semantic metrics described previously. Starting with researchers (~1400) who were affiliated with The Ohio State University, paths with 5 keywords were generated to link them with researchers from outside this group (~14000). In other words, the generated paths linked the researchers from The Ohio State University with researchers from other institutions, five keywords apart. Paths generated for each researcher were grouped into different files. An example of the generated paths is depicted in Figure 4. Note the diamonds in red in Figure 4 that indicate the starting and ending keywords of the path. A researcher (anonymized) using the keyword *"Grassland"* is connected by the path to a researcher using the keyword *"Hyperplasia."*

Once these paths were generated for each researcher affiliated with The Ohio State University (OSU), an association matrix M_{assoc} was created where each row contained the researchers x_i connected with an OSU researcher o_i through some path (6).

$$M_{assoc} \equiv \begin{bmatrix} o_1 & x_1 & \cdots \\ \cdots & \cdots & \cdots \\ o_n & x_i & \cdots \end{bmatrix} \quad (6)$$

From this matrix, clusters of OSU researchers were derived such that each cluster contained OSU researchers and the largest intersection of associated external researchers. The clustering was performed with a threshold value of 10 that corresponded to the minimum cardinality (number of researchers) of the intersection set. This was the final output of the knowledge discovery methodology; clusters of researchers that could potentially collaborate on cross-disciplinary, inter-institutional research projects along with a numeric estimate of their viability viz., the path weight metric.

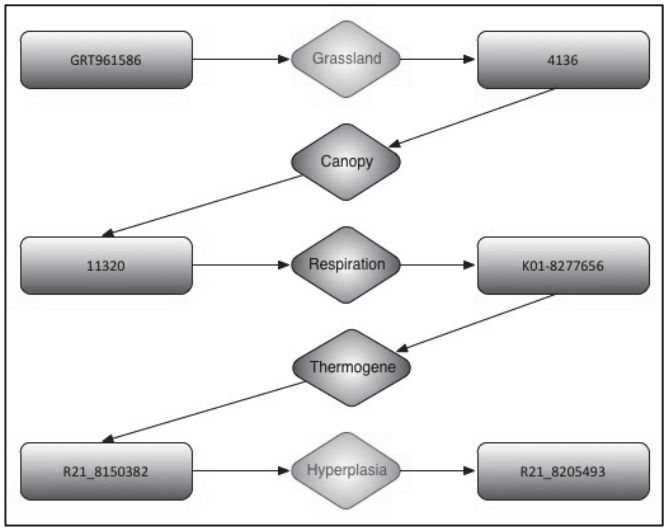


Figure 4: An example of a generated path

4. Discussion

In the previous sections, we have described an intelligent, semantically enriched knowledge discovery algorithm that extracted groups of collaborating researchers from different disciplines and institutions. Starting with a set of keywords and the researchers they were associated with (and vice versa), the algorithm derived sets of collaborators. Further, the results were filtered and ranked based upon a quantitative semantic metric, the Path Weight Threshold. The entire process is represented in Figure 5; starting from a collection of keywords that are processed systematically until the final step, which outputs sets of collaborating researchers. The path finding algorithm is different from classical clustering techniques in that it brings together researchers from widely differing backgrounds; clustering techniques would typically group researchers from similar backgrounds.

The algorithm relies on simple keywords to connect pairs of authors; a possible limitation. The classification of keywords

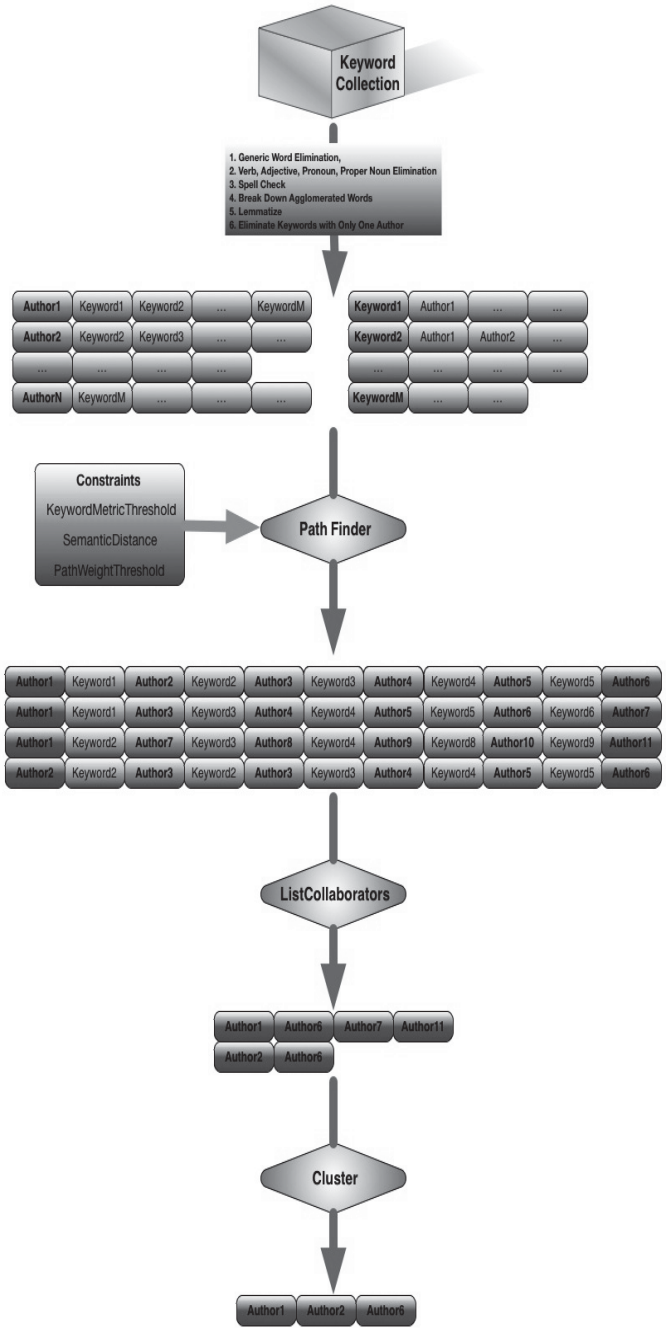


Figure 5: Overview of the knowledge discovery methodology

to specific subheadings and the addition of metrics addresses this limitation. However, the classifications themselves are subjective derived from manual annotation. The use of crowdsourcing techniques where annotations from volunteers can be summarized [10] may address this. Upper-level ontologies such as the Basic Formal Ontology (BFO, www.ifomis.org/bfo/) and Cyc (www.cyc.com/platform/opencyc) may also be used to classify the keywords. OpenCyc provides an API and a knowledge base browser that can be used in the keyword classification and metric computation process. Lastly, the identity of the associated researcher could be very useful in obtaining the correct context of context-sensitive keywords, given all the researchers have been de-identified in the input data set.

The semantic metrics computed for each keyword take into account the specificity of the subheading under which it is classified and also the extent of its usage. A limitation of this approach is that keywords that are classified into shallower subject heading hierarchies (such as *Psychology*) may be weighted lower than those that are classified into deeper hierarchies. We have attempted to minimize this effect by restricting the depth of each hierarchy to three levels. Moreover, very few keywords (less than 0.5%) are classified into the shallower hierarchies with 2 levels or less.

The subject heading hierarchies used here are based upon the Library of Congress classification scheme. The use of other hierarchies (such as OpenCyc) in the future will create a parallel result set for validation and summarization.

5. Conclusions

In this paper, we have presented a semantically enriched knowledge discovery methodology, which generates paths through a hyperconnected graph comprising researchers as nodes and associated keywords as connecting edges. The generated paths connect researchers from differing areas of expertise creating groups of potential collaborators, who can identify and work on pressing translational research problems. This methodology can be readily implemented to establish connections between other entities as well, for example, in drug repurposing [11] and in hypothesis generation for connecting phenotype and genotype data [12].

6. Acknowledgments

This work was supported by Award Number P01CA081534 from the National Cancer Institute of the National Institutes of Health. I would like to acknowledge Philp Payne, Jason Sullivan, and Julia Carpenter-Hubin for their role in data collection and preliminary data analysis.

7. References

[1] Woolf SH. The meaning of translational research and why it matters. JAMA 2008: 299(2): 211 – 213. doi:10.1001/jama.2007.26

[2] Rubio DM, Schoenbaum EE, Lee LS, et al. Defining translational research: Implications for training. Academic Medicine 2010: 85(3): 470 – 475. doi: 10.1097/ACM.0b013e3181ccd618

[3] Box-Steffensmeir J, Hoy C, Martin W et al. Summary report on Development of a singular presence in Data Analytics for The Ohio State University. Technical Report. 2014.

[4] Manning C, Raghavan P, and Schutze H. Hierarchical Clustering. In: Introduction to Information Retrieval (Chapter 17). Cambridge University Press. 2008. Pp. 377 – 401.

[5] Manning C, Surdeanu M, Bauer J et al. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proc. of the 52nd Annual Meeting of the Assoc. of Comp. Linguistics: System Demonstrations, 2014; pp. 55 – 60.

[6] Norvig P. How to write a spelling corrector. http://norvig.com/spell-correct.html

[7] McGuinness DL and van Harmelen F. OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/

[8] Gruber T. A translational approach to portable ontologies. Knowledge acquisition 1993: 5(2): 199 – 220.

[9] Sedgwick R and Wayne K. Searching. In: Algorithms (Chapter 3). Addison Wesley. 2011.

[10] Good BM and Wilkinson MD. Ontology engineering using volunteer labor. In: Proc. of the 16th Int'l World Wide Web Conf. (WWW 2007). Banff, Canada. Pp: 1243 – 1244.

[11] Regan K, Raje S, Kothari C, and Payne P. Conceptual knowledge discovery in databases for drug combination predications in melanom. In : Proc. of the 2015 AMIA Joint Summits on Trans. Science, San Francisco CA. March 2015

[12] Payne P, Borlawski T, Lele O et al. The TOKEn project: knowledge synthesis for in silico science. JAMIA 2011: 18:i125 – i131. Doi:10.1136/amiajnl-2011-000434

Condensation of Reverse Engineered UML Diagrams by Using the Semantic Web Technologies

Meisam Booshehri, Peter Luksch
Institute of Computer Science, University of Rostock, Germany

Abstract - *Up-to-date software design documentation is valuable for maintenance engineers, testers and developers joining a project at a later stage; however, UML Models, e.g. class diagrams, are often poorly kept up-to-date during development and maintenance. Reverse engineering, therefore, has become a popular method to recover an up-to-date design from the underlying source code. However current techniques yet produce a detailed representation of the underlying source code that would reduce the understandability. In order for the understandability to enhance, condensation of the reverse engineered diagrams has been proposed as a solution. However, current state-of-the-art approaches in this area still demand a need for improving or alternative approaches. Consistently, in this paper we are putting forward a bridging idea of using the semantic web technologies for improving the condensation process. Two contributions are proposed that support each other. Firstly, the V-Ontmodel is suggested for updating software documentation over the software evolution. Secondly, a general architecture is proposed enabling us to reduce sophisticated analysis tasks for the condensation process of class diagrams to a few queries in SPARQL or its extensions like SPARQL-ML. To discuss the feasibility of the approach we focus on the condensation of reverse engineered class diagrams in the whole paper. Finally, an illustration example is presented in which helper classes are excluded from the class diagrams in the condensation process regarding the structural patterns extracted from the software metadata.*

Keywords: *reverse engineering; simplification of UML diagrams; semantic web technologies; up-to-date design; condensation of class diagrams*

1 Introduction

Booch et al. proposed the Unified Modeling Language (UML) as a standard for writing the software blueprints. "UML may be used to visualize, specify, construct and document the artifacts of a software-intensive system"[1]. Currently UML 2.0 is an ISO standard where it provides 13 different diagrams so that one can express all the important features of a system. One of the UML diagrams is the Class Diagram by which developers could model classes in an object oriented design including their attributes, operations and their relationships and associations with other classes.

In practice as software evolves, design documentations will become more and more outdated. UML models are often poorly kept up-to-date during development and maintenance. On the other hand, an up-to-date design is of special importance for maintenance engineers, testers, developers, and other engineers joining a software project at a later stage. As a solution, consequently, reverse engineering methods have been proposed in order to recover up-to-date design diagrams automatically[2].

However, current reverse engineering techniques do not yet solve the problem sufficiently. Particularly, the Computer-aided Software Engineering (CASE) tools that offer functionality for reverse engineering into UML class diagrams produce a detailed representation of the underlying source code that would reduce the understandability[3]. Therefore, software engineers would come across difficulties in recognizing the key elements in a software structure.

In order to the increase understandability, Osman et al. [3-5] propose the condensation of reverse engineered class diagrams where the elemental question is how to select the key elements in the software architecture to recover class diagrams of suitable abstraction level? Subsequently, Thung et al. [6] propose an extension to Osman et al.'s work which brings an improvement to the previous approach. However, according to Thung et al., there are threats to the validity of their approach, specifically for the generalization of their approach to other projects. Moreover, the quality of reverse engineered diagrams has yet to be improved in order to become more and more similar to their forward design counterparts in terms of understandability.

Consistently, it is our goal in this paper to explore ways in order to facilitate current methods of condensation of reverse engineered diagrams. Since software engineering is naturally a knowledge-intensive activity, we are motivated to put forward a general bridging idea of using the semantic web technologies for improving and facilitating the process of condensation of reverse engineered UML diagrams. We make use of the semantic web technologies such as RDF and OWL in order to bring the ability of knowledge management for the machines and facilitate the condensation process.

Overall, the hypothesis we are going to verify is the following: *"we can use the semantic web technologies in order to provide a semantic infrastructure for storing software metadata and reasoning about them enabling an easier recovery of UML diagrams from the underlying source code where sophisticated condensation analysis tasks will be reduced to some few SPARQL queries."*

The rest of the paper is organized as follows. The second section is to review the background and some popular related

work. Then, in the third section the proposed approach is described and a general architecture is presented in order to show the feasibility of the approach. The fourth section will then come up with an illustration example. Finally, the fifth section is to conclude and discuss the future work.

2 Related Work

The related work to our approach is divided into two categories. The first category is related to the applications of the semantic technologies in the software development life cycle and the second category includes the current methods for condensation of reverse engineered diagrams. Therefore, in the next subsections we review some popular work in these two areas.

2.1 The semantic technologies and software engineering

The semantic web technologies such as RDF and OWL bring the ability of knowledge management for the machines, and consequently, they can help software engineers with automating different activities involved in software development life cycle. There are a significant number of research activities in this field which utilize the semantic web technologies including requirement engineering [7, 8], software testing[9], automatic component selection [10], automatic model creation from textual specifications of the softwares[11-13], and automatic detection and selection of design patterns[14-16]. The most related approach to our paper is the Tappolet et al.'s approach where they propose an approach for reasoning about software evolution[17]. They introduce EvoOnt Ontologies which provide the basis for representing source code and meta-data in OWL. This representation reduces analysis tasks to simple queries in SPARQL (or its extensions). There is also a limitation to their approach which is the loss of some information due to the use of the FAMIX meta model. The FAMIX describes the core structure of object oriented software without being fixed to one programming language such as C++, Java, etc; however, it does not model language constructs like switch-statements. As a result, by using the FAMIX model measurements cannot be conducted at the level of statements.

2.2 Condensation of Reverse Engineered UML Diagrams

As for the condensation of the reverse engineered class diagrams, according to Osman et al. [4], *GUI-related Information*, *Private and Protected Operations of a class* and *Helper classes* can be excluded from a class diagram in order to increase the understandability. There are also another measures that can be considered in the condensation process. For instance, it is a common belief that the classes which frequently changes over software evolution are considered as candidates to key classes in a class diagram. Besides, according to Osman et al.[5] the number of public operations of a class is one the most important metrics indicating the importance of the class.

3 Proposed Approach

Our main idea is to make use of the semantic web technologies in order to prepare a suitable infrastructure for reasoning about the evolution of a software project enabling a more accurate recovery of UML diagrams from the underlying source code. As a result we expect to become capable of performing sophisticated condensation analysis with only some few SPARQL queries. By providing a framework for representing software source code and meta-data in an OWL ontology, we will be able to reason over software meta-data and thus facilitating software understandability and maintenance

In this section therefore, we introduce a general model for updating software documentation over the software evolution by using the ontology-based modeling. Next, we deepen our approach by focusing on the reverse engineered class diagrams and a semantic model is proposed specifically for the condensation of the reverse engineered class diagrams.

3.1 The V-OntModel for Updating Software Documentation

According to the history of utilizing the semantic technologies in different steps of software development life cycle as discussed in the related work section, we propose a general model called the *V-OntModel* for updating software documentation at different stages of software development by using ontology-based modeling of the software. We can argue that the main benefit of using the V-OntModel is to help overcome lack of information about the elements of the artifacts produced in the software evolution. For instance, consider a class diagram and a metric called the Number of Revisions (NOR) of a class which enables us to recognize the key classes in an application. The more revisions made to a class, the more important the class is. By exploiting the V-OntModel from the beginning of the development, we will be able to record the number of revisions of a class while this is not possible when we do not store metadata about a software over its evolution.

The main idea behind the V-OntModel is that the ontologies can define any kind of relationships between different items in the micro real world at different levels of abstraction, thus various types of UML diagrams can be modeled by using ontologies at different levels of abstraction.

As represented in Figure 1, the V-OntModel depicts the relationship of ontology-based modeling actions to the actions associated with requirements engineering, software design and construction activities. As a software team moves down the left side of the V, basic problem requirements in the micro real world are refined into more and more technical representation of the problem while also ontology-based models are being created (semi)automatically. Once the

software code has been generated, we can move up the right side of the V, making use of ontology-based models, essentially performing a series of reverse engineering tasks to recover up-to-date design documentations. The V-OntModel provides a way of visualizing how reverse engineering tasks are applied to different software development steps by creating the ontology-based models.

In order to illustrate how the V-OntModel works, we draw your attentions to two examples regarding *Component Diagrams* and *Use-Case Diagrams*. In a semantic infrastructure, changes to object oriented code or requirements analysis documents can be controlled by a *software ontology* Such as EvoOnt [17] and directly applied to UML diagrams.

- Example 1: As for a Component Diagram, in the case of any change in the relationships between components in the source code, the changes can be applied simply to the Component Diagram with respect to the software ontology.
- Example 2: Regarding the Use-Case Diagrams, any considerable change in user stories prepared in the requirement analysis can be applied to the software ontology, thus changing the use-case diagram.

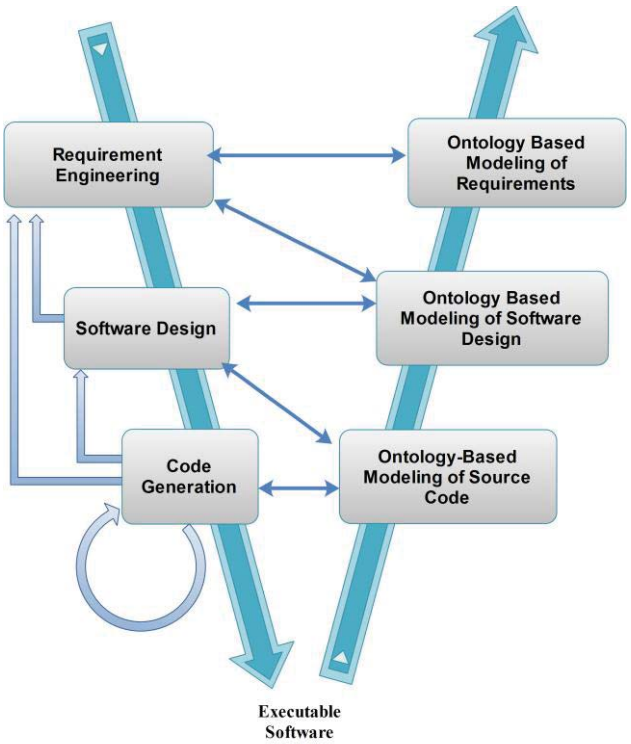


Figure 1 The V-OntModel

3.2 Proposed Architecture

One of the suitable architectures that can show the feasibility of our approach has been depicted in Figure 2. It consists of seven components: *Source code Parser*, *RDF/OWL Converter*, *Reference Ontology*, *Decision List*, *SPARQL Query Generator* and *Condensation Unit*. We discuss these components in more details in the sections below.

3.2.1 Source Code Parser

To parse the source code and generate a first model of the software project, a Source Code Parser has been embedded in the architecture in order to recognize different components like Class, Method or Attribute in an Object Oriented source code. Besides, the model is expected to contain different relations between the components such as association relations.

3.2.2 RDF/OWL Converter

The model created by the Source Code Parser has to be converted into RDF/OWL format in order to populate the reference ontology. Reference Ontology is a manually designed OWL ontology enabling us to store meta-data about a software system and reason over them. Naturally, modeling system metadata in OWL brings us Interchangeability, Non-ambiguity, Machine-readability and Extensibility[17].

By embedding a RDF/OWL converter in the proposed architecture, a database of RDF triples is produced as the software metadata repository.

3.2.3 Condensation Unit

The Condensation unit consists of a decision list and a SPARQL Query Generator. In this unit, the RDF data produced by the RDF/OWL converter is analyzed so that we can rank the classes and then decide on the candidate classes to be excluded or included in the reverse engineered class diagram according to a decision list as will be discussed. Obviously, there should be a SPARQL Query Generator embedded in the architecture in order to extract the insight from the provided database.

In the following section we elaborate on the decision list and the SPARQL Query Generator.

3.2.3.1 Decision List

There are a number of measures that can be considered for condensing a reverse engineered class diagram. For instance, it is believed that the classes which frequently changes over software evolution are considered as candidates to key classes. In addition, Osman et al. [4] provide a research which concludes the information that should be excluded from a class diagram includes *GUI-related Information*, *Private and Protected Operations of a class* and *Helper classes*.

Besides, according to Osman et al.[5] the number of public operations of a class is one of the most important metrics indicating the importance of the class. Accordingly, we believe that different kinds of *cohesion* and *coupling* are important in order to recognize the key classes in a component-level design. As for the coupling, software

engineers try to keep it as low as possible in order to decrease the complexity of the system; however, sometimes high coupling cannot be avoided and its ramifications have to be understood[1]. Consequently, when a class is inevitably highly coupled with other classes through a large number of public operations, it has to be a sign of importance for that class.

Overall, there should be a list of measures for the proposed architecture in order to recognize the key classes. We propose a classification of the most important factors into three categories:

- **Software Design Patterns:** Software design patterns might include classes which do not play a top priority role in the application. The proposed procedure is therefore to mine the software metadata repositories for finding different design patterns by posing few SPARQL Queries. Next, we will be able to recognize the low priority classes in a design pattern in order to exclude them from reverse engineered class diagrams. To make this subject clear, an illustration example is proposed in the next section.

- **Metrics:** Tappolet et al. [17] compute different metrics based on their reference ontology that can be useful for our objectives as well. Among these metrics, NOPA (Number of Public Attributes) and NOR (Number of Revisions) can be of valuable help for deciding on the key classes.

- **Algorithms:** while deciding on the key classes by using the machine learning techniques, we have to decide on different methods such as Relational Probability Trees and Relational Bayesian Classifier. The end user can make a final decision on this task.

3.2.3.2 SPARQL Query Generator:

In the proposed architecture, the SPARQL Query Generator will pose SPARQL queries based on the decision list where it might include machine learning tasks. Therefore, we can make use of SPARQL extensions such as SPARQL-ML or any other customized version of SPARQL.

4 Illustration Example

A **helper class** is used in object oriented programming in order to assist some functionality, which is not the primary objective of the application or class in which it is invoked. An instance of a helper class is called a **helper object**. As correctly argued by Osman et al. [4] helper classes should be excluded to simplify a reverse engineered class diagram. In order for that to happen, our hypothesis is to mine the metadata repositories for finding structural patterns such as delegation patterns in which a helper object can be occurred. A **delegation pattern** in software engineering is a design pattern that delegates a responsibility to an associated helper object. Fortunately, finding each type of a structural pattern, such as delegation pattern can be done by posing few SPARQL queries over the software metadata repository[17]. Finally, after finding delegation patterns and helper classes inside them, we can decide on each of helper classes to be excluded or included based upon, for instance, user-defined conditions in a (semi)automatic mode.

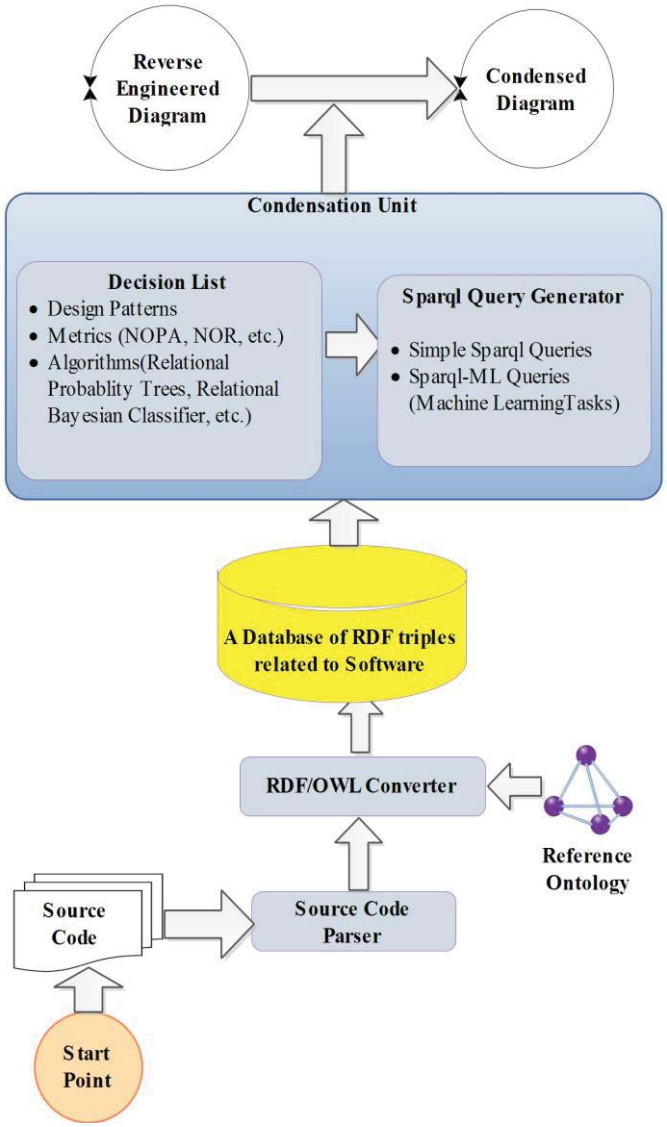


Figure 2 The Proposed Architecture

5 Conclusion and Future Work

In this paper, we are investigating effective methods to (semi)automatically recover an up-to-date UML design documentation which is helpful for the software engineers joining a project at a later stage. Reverse engineering is a popular method for this goal; however, current techniques provide detailed reverse engineered UML diagrams which are difficult to understand. Therefore, we propose two

contributions while the focus of this research remains on the class diagrams. First, the V-OntModel is proposed to help overcome lack of information about the elements of the artifacts produced in the software evolution, such as classes in a class diagram. Next, in order to reach a higher level of abstraction and increase the comprehensibility of reverse engineered class diagrams, a condensation architecture is proposed in which the classes are ranked according to a decision list. Then, a final decision can be made on the candidate classes to be excluded from a reverse engineered class diagram. Finally, we can make an obvious argument as follows. When the "V-OntModel" is used over the software evolution, the proposed condensation architecture can achieve better performance in reverse engineering.

As for the future work, we will implement an open-source core system to which everybody can add his own packages for handling different kinds of condensation while there will be also the possibility of adding different algorithms to the core system for a specific goal.

6 References

[1] R. Pressman, *Software Engineering: A Practitioner's Approach*: McGraw-Hill, Inc., 2010.

[2] H. Osman and M. R. Chaudron, "Correctness and Completeness of CASE Tools in Reverse Engineering Source Code into UML Mode," *GSTF Journal on Computing,* pp. 193-201, 2012.

[3] H. Osman, *et al.*, "An Analysis of Machine Learning Algorithms for Condensing Reverse Engineered Class Diagrams," unpublished.

[4] H. Osman, *et al.*, "UML class diagram simplification: what is in the developer's mind?," in *Proceedings of the Second Edition of the International Workshop on Experiences and Empirical Studies in Software Modelling*, 2012, p. 5.

[5] H. Osman, *et al.*, "UML Class Diagram Simplification - A Survey for Improving Reverse Engineered Class Diagram Comprehension," in *1st International Conference on Model-Driven Engineering and Software Development*, 2013, pp. 291-296.

[6] F. Thung, *et al.*, "Condensing class diagrams by analyzing design and network metrics using optimistic classification," in *Proceedings of the 22nd International Conference on Program Comprehension*, 2014, pp. 110-121.

[7] S. J. Körner and T. Brumm, "Improving Natural Language Specifications with Ontologies," in *SEKE*, 2009, pp. 552-557.

[8] M. Ilieva and H. Boley, "Representing Textual Requirements as Graphical Natural Language for UML Diagram Generation," in *SEKE*, 2008, pp. 478-483.

[9] S. Paydar and M. Kahani, "Ontology-based web application testing," in *Novel Algorithms and Techniques in Telecommunications and Networking*: Springer, 2010, pp. 23-27.

[10] O. Hartig, *et al.*, "Automatic component selection with semantic technologies," in *Proc. s of the 4th Int. Workshop on Semantic Web Enabled Software Engineering (SWESE) at ISWC*, 2008.

[11] A. Fatolahi, *et al.*, " A Model-Driven Approach for the Semi-Automated Generation of Web-based Applications from Requirements," in *2008 International Conference on Software Engineering and Knowledge Engineering*, 2010, pp. 619-624.

[12] Y. Tian, *et al.*, "An Ontology-based Model Driven Approach for a Music Learning System," in *SEKE*, 2009, pp. 739-744.

[13] S. J. Körner and T. Gelhausen, "Improving Automatic Model Creation Using Ontologies," in *SEKE*, 2008, pp. 691-696.

[14] J. Dietrich and C. Elgar, "A formal description of design patterns using OWL," in *Software Engineering Conference, 2005. Proceedings. 2005 Australian*, 2005, pp. 243-250.

[15] J. Dietrich and C. Elgar, "An ontology based representation of software design patterns," *Design Patterns Formalization Techniques,* p. 258, 2007.

[16] H. Kampffmeyer and S. Zschaler, "Finding the pattern you need: The design pattern intent ontology," in *Model Driven Engineering Languages and Systems*: Springer, 2007, pp. 211-225.

[17] J. Tappolet, *et al.*, "Semantic web enabled software analysis," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 8, pp. 225-240, 2010.

Determinants of Plasma Beta-Carotene Levels Using Neurofuzzy System and Multivariate Analyses

Deok Hee Nam
Engineering and Computing Science, Wilberforce University, Wilberforce, OHIO, USA

Abstract - *The observational study of the determinants of Plasma Beta-Carotene Levels through the investigation of the relationship between personal characteristics and plasma concentrations of beta-carotene has been examined by the various multivariate analysis techniques using the neurofuzzy systems with applying the various reduced data sets without losing any significant meaning comparing to the original data. In addition, various statistical categories to evaluate the best-fitting model using the original data as well as the reduced dimensional data. In order to reduce the dimensionality of the original data, various multivariate analysis techniques are applied with different types of the statistical distributed methods.*

Keywords: data mining; dimensional reduction; multivariate analysis; and neurofuzzy system

1 Introduction

The studies about beta-carotene and retinol are most extensively examined components in various populations, for both human plasma concentrations and dietary intake because of their inverse relationships with the development of several diseases such as cancer, cardiovascular disease, and cataracts. In addition, it is difficult to determine the levels of the plasma beta-carotene using various conditions. Simultaneously, in order to optimize the decision of the plasma beta-carotene levels using personal characteristics, there are various statistical techniques to handle the relatively larger data system more efficiently with a smaller dimensional system to identify or implement the original data system. Even though many scientists are presenting various techniques in these days, it is not simple to reduce a relatively higher number of variable-data into a smaller number of the variable-data efficiently and equivalently without losing any significant meaning of the original data set. Moreover, it is even more difficult to find the best-fitting model or a technique with an appropriate process and interpret the original higher variable-data system. Among the various statistical methods with optimizing techniques of the data reduction, the ideal technique is to diminish the number of variables from the original data into the smaller number of variables with the reduced number of observed instances simultaneously. To accomplish the reduction of the higher variable number system, the multivariate analysis methods such as principal component analysis, factor analysis, or clustering analysis are frequently used. These multivariate analysis techniques are generally seeking the underlying structure by extracting uncorrelated components or factors without losing any significant meaning of the original data after eliminating the redundancies of the original data set. In the paper, the hybrid procedures of multivariate analysis techniques with various transforming techniques (like orthogonal transformation, varimax rotation, or etc.) are examined. In addition, the maximum likelihood evaluation is used along with different types of statistical methods, multivariate analyses including principal component analysis and factor analysis. Those applied techniques recognize a reduced set of grouped observations and a set of uncorrelated variables of the given data. Those reduced sets are employed by the various neurofuzzy systems as a form of the underlying structure. The developed neurofuzzy systems with the reduced sets assess the predicted values to evaluate the extracted sets to verify how closely the reduced sets can interpret the original data set. Moreover, since the examined data system cannot be expressed by a certain mathematical expression, neurofuzzy systems are used to evaluate the decision of the plasma concentration level using the beta carotene without using any particular mathematical equations. The section 2 describes the related techniques and the section 3 is about the employed data set followed by the neurofuzzy implementation in the section 4. In the section 5, analyses and results with statistical categories are mentioned. The section 6 is about the conclusion using the proposed techniques with the beta carotene levels.

2 Review of literature

2.1 Principal Component Analysis

In general, Principal Component Analysis (PCA) [1, 12] measures directly observed variables, which form a linear combination of weighted observed variables with uncorrelated and orthogonal from the given data. This technique is one of the most popular techniques along with the factor analysis to identify the underlying structure from the given data with the reduction of the dimensionality for the original multi-variables data systems. To apply principal component analysis (PCA), the total variance of the reduced or extracted components are calculated in order to construct a newly reduced multi-variable data systems without losing any significant meaning of the original data and eliminating highly correlated components from the original variables. Sainani [1] and Jolliffe [11] comprehensively presented all required procedures of PCA. The following steps are briefly describing how to accomplish the steps of PCA procedures to extract the required principal components from the original data. Let X be an n dimensional data set with $m \times n$ matrix

format, i.e. X = {$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …….$\boldsymbol{x}_n$}, where n is the number of measurement variables to represent the dimensionality of the given multidimensional data system and m is the number of observed instances for the data. First, standardize X by normalizing $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …….$\boldsymbol{x}_n$, with subtracting the mean from each measurement variable, respectively. After the standardization of X, apply Singular Value Decomposition (SVD) technique with calculating the eigenvalues and the eigenvectors of the covariance (or correlation) for the newly extracted components. Finally, calculate the component coefficients and the component scores in order to find out the newly reduced principal components through determining the less dimensionality of X with most meaningful components by accumulating the calculated covariance based upon the required criterion.

2.2 Factor analysis

Factor analysis (FA) [2, 12] is a statistical technique that examining and extracting the embedded factors through underlying latent variables which represent the original variables. These factors are mostly orthogonal and ordered by the proportion of the variance of the original data. Among the original variables y_1, y_2, ... , y_p, with the moderately correlated factors, there is a possibility to reduce the basic dimensionality of the original data into a less dimension, q, with q < p. In general, factor analysis used only a subset of identified factors from the original data for recognizing the structure of the original data and the remaining factors are reflected as irrelevant or not necessary for the identification of the original data structure. In other words, when factor analysis is compared with principal component analysis, factor analysis is to recognize the grouped scheme of the original data structure with the less dimensionality applying newly extracted factors after eliminating the redundancies of the original variables. The procedure of the factor analysis can be expressed as the following steps. Based upon the given multi-variable data in a vector form, F, such as

$$F = \{F_1, F_2, \dots, F_N\} = \left\{ \begin{bmatrix} f_{11} \\ \vdots \\ f_{m1} \end{bmatrix} \dots \begin{bmatrix} f_{1n} \\ \vdots \\ f_{mn} \end{bmatrix} \right\}. \quad (1)$$

where m is the number of the observed instances and n is the number of the vector, F_i. Using F from the equation (1), first, the correlations of the response vector F, denoted by the correlation matrix, Σ, is calculated by evaluating the correlation coefficients between F_i. Then, the standardization of the original variables with the normalization of each variable is applied in order to construct the correlation matrix using the basic input factors to a common factor analysis. Using the correlation matrix, the eigenvalue, λ, using the characteristic equation is calculated. Applying the eigenvalues, if $i = j$ where i and j are the indices of the correlation matrix Σ, then, the diagonal elements of the matrix Σ are $\sqrt{\lambda_i}$, and 0 if $i \neq j$. Then, initial factor loadings are calculated are created and followed by the Varimax rotation (if the rotation is applied) with applying the rotated factor loadings. Finally, newly extracted factor scores can be calculated as the projection of an observation on the common factors.

2.3 Maximum-likelihood estimation

The maximum-likelihood estimation (MLE) [3, 4] is the procedure of finding the value of one or more parameters of a statistical model which developed by R.A. Fisher in the 1920s. When applied to a data set and given a statistical model, the maximum-likelihood estimation provides the estimation for the model's parameters.

Let X_1, . . . , X_n be n independent random variables with probability density functions (pdf) $f_i(x_i ; \theta)$ depending on a vector-valued parameter θ, which identifies the data-generating process that underlies an observed sample of data. The pdf provides a mathematical description of the data which is the product of the individual densities. This joint density is called "likelihood function", which defined as a function of the unknown parameter vector, θ, where y is used to indicate the collection of sample data. Consider the random sample with a certain conditions of the observations. If the joint density gives that the value of θ can make the sample data, x, this estimation is called as maximum likelihood estimate, or MLE, of the unknown parameter, θ. The maximum likelihood estimate of θ, $\hat{\theta}_{\text{MLE}}$,[15] can be defined by the value of θ in parameter space Ω that maximizes the likelihood function, $L(\theta/x)$ [15] where

$$\hat{\theta}_{MLE} = \underset{\theta \in \Omega}{max} L(\theta/x) = \underset{\theta \in \Omega}{max} \prod_{i=1}^{n} f(x_i/\theta) \quad (2)$$

and

$$L(\theta/x) \propto f(x/\theta) = \prod_{i=1}^{n} f(x_i/\theta). \quad (3)$$

2.4 Varimax rotation

The varimax rotation [13] was developed by Kaiser (1958) [5] and used as the most popular rotation method for factor analysis. The varimax rotation is a technique to rotate the orthogonal basis to align with the related coordinates in order to simplify the interpretation of the particular sub-space without changing the actual coordinate system. The procedure of the varimax rotation involves scaling the loadings by dividing them by the corresponding communality as

$$\tilde{\iota}_{ij}^{\,*} = \frac{\hat{\iota}_{ij}^{\,*}}{\hat{h}_j} \quad (4)$$

The loading of the i^{th} variable on the j^{th} factor after rotation, where $\hat{h}_j$ is the communality for variable i. To maximize the quantity, the varimax procedure selects to maximize the variances of the standardized loadings for each factor with the summation over the m factors such as

$$\text{V} = \frac{1}{p}\sum_{j=1}^{m}\left\{\sum_{i=1}^{p}\left(\tilde{\iota}_{ij}^{\,*}\right)^4 - \frac{1}{p}\left(\sum_{i=1}^{p}\left(\tilde{\iota}_{ij}^{\,*}\right)^2\right)^2\right\} \quad (5)$$

After applying a varimax rotation, each original variable is closely associated with one (or a small number) of extracted factors, and each factor represents only a small number of variables. In general, the varimax rotation searches for a rotation (i.e., a linear combination) of the original factors such that the variance of the loadings is maximized

2.5 Neurofuzzy system

A neurofuzzy system [6, 14] is a hybrid system which is a fuzzy logic system based upon fuzzy if-then rules with appropriate membership functions to generate input output pairs applied by neural network technique that uses a learning algorithm derived from examined and trained data to determine the developed system's characteristics. Jang introduced Adaptive Neuro-Fuzzy Inference System (ANFIS)[7], which represents a structure of a neurofuzzy system based upon Takagi–Sugeno fuzzy inference system using five different layers such as input layer, production layer (fuzzification), normalized firing layer (inference), consequence parameters layer (defuzzification), and finalized output layer. Fig. 1 shows the structure of ANFIS system with five network layers.

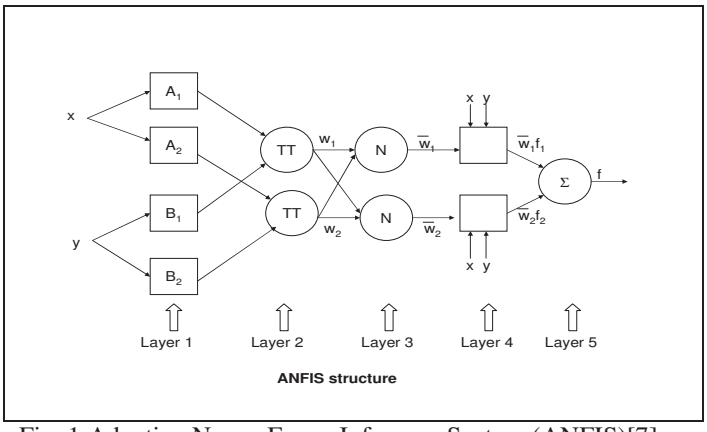


Fig. 1 Adaptive Neuro-Fuzzy Inference System (ANFIS)[7]

Layer 1 is known as the fuzzification layer with the membership function, O_k^1, where $O_k^1 = \mu A_k(x_k)$, x_k is the input, and A_k are the linguistic labels associated with the membership function. Layer 2 is the multiplication of input signals to produce the degrees, w_i, of the membership function. Layer 3 is the normalization layer associated with the rule's firing strengths. Layer 4 is the defuzzification layer as the weighted consequent value, $\overline{w} f_k$. The output from Layer 5 from Fig. 1 is the summation layer for ANFIS [14] expressed by

$$O_{5,i} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad (6)$$

with applying the following inference rules [7] for the neurofuzzy systems;

if x is A_1 and y is B_1, then, $f_1 = p_1x + q_1y + r_1$,
or
if x is A_2 and y is B_2, then, $f_2 = p_2x + q_2y + r_2$.

3 Data of plasma concentrations [8]

The original data set consists of low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids which might be associated with increased risk of developing certain types of cancer. Since the data describes a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. The applied data in this paper used only the data set of the female patients from the total study subjects (N = 315), which were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. Among the female patients data, only 7 out of 14 measurements types are selected such as quetelet (weight/square of height), calories (number of calories consumed per day), fat (grams of fat consumed per day), fiber (grams of fiber consumed per day), alcohol (number of alcoholic drinks consumed per week), cholesterol (cholesterol consumed milligrams per day), and beta-plasma (plasma beta-carotene ng/ml).

4 Applied neurofuzzy systems

There are seven different neurofuzzy systems using Adaptive-Network-Based Fuzzy Inference Systems (ANFIS) to predict the determinants of Plasma Beta-Carotene Levels using the less dimensionality of the selected variables applying various statistical algorithms such as principal component analysis (PCA), and factor analysis with and without varimax rotation or maximum likelihood estimation from the six original measurements types and five reduced measurements types. The following figures are about the generated neurofuzzy system with the five reduced measurements types. Fig. 2 shows the properties of neurofuzzy system with five inputs and one output. As shown in Fig. 3, Gaussian membership functions are used for the system membership functions for each input and output for the neurofuzzy systems. Fig. 4 shows the applied rules for the neurofuzzy system and Fig. 5 describes how the structure of the neurofuzzy system is developed based upon ANFIS. There are five layers to extract the finalized output through fuzzification and defuzzification procedures as shown in layer 2 to layer 4 from Fig. 5. In Fig. 6, the surface plot is presented by visualizing the plasma concentration level data that are larger to display in numerical form and for graphing functions for the multi-dimensionalities of the plasma concentration level data.

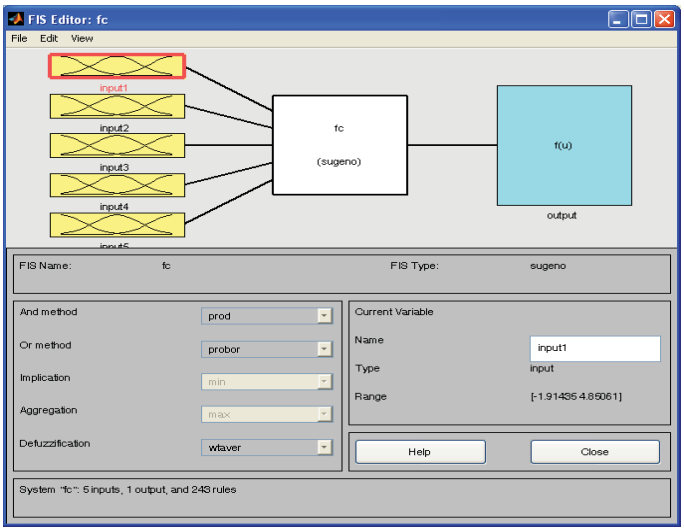


Fig. 2 Neurofuzzy inference system with properties including three inputs and an output.

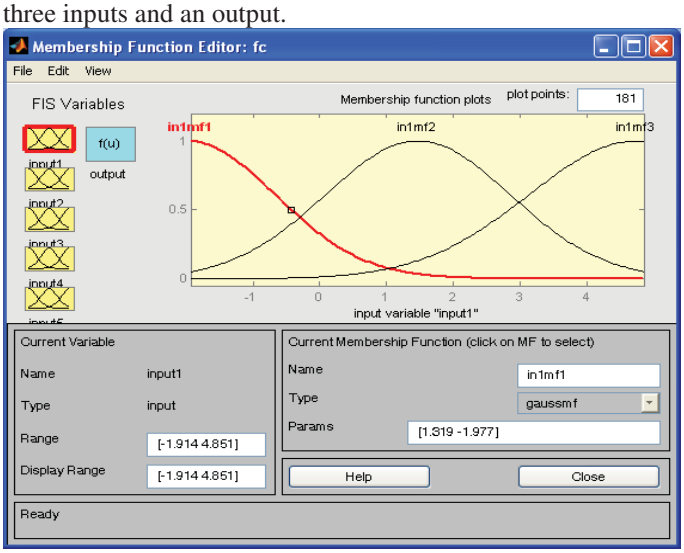


Fig. 3 Neurofuzzy inference system with membership functions

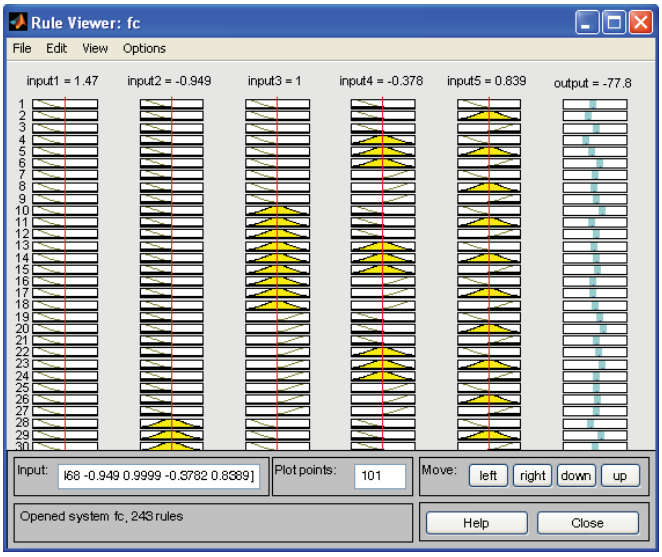


Fig. 4 Neurofuzzy inference system with applied rules for defuzzification

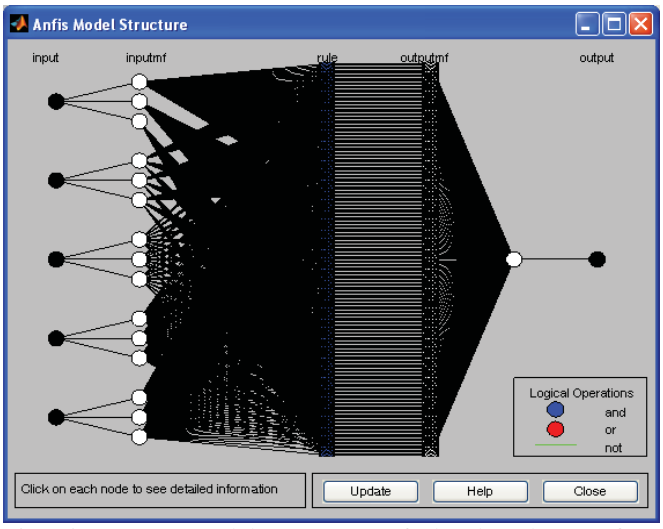


Fig. 5 ANFIS Model Sturcutre of developed neurofuzzy inference system with five inputs

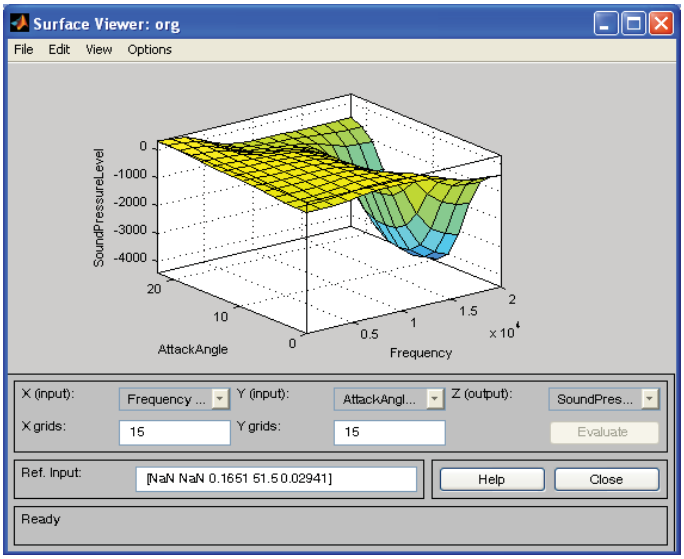


Fig. 6 Surface viewer to display in the numerical form of the plasma concentration data set

5 Analysis and results

To determine the level of the plasma concentration for the beta-carotene, six inputs and one output are originally applied. As shown in Fig. 7, all possible eigenvalues are presented based upon the obtained variances of the newly extracted components from the five original measurements types. "The Eigenvalues-Greater-Than-One Rule" [9], and 0.9 or above criterion for the accumulation of the variances are used to predict the level of the beta carotene based upon the generated neurofuzzy systems using the five reduced components from the six original variables determined by the accumulation of the variances from the newly extracted components. Five different statistical methods are compared with the predicted values using the various neurofuzzy systems applying seven different plasma concentration data sets; ORG, FC, FCV, FM, FMV, PCR and PCV. ORG stands

for the neurofuzzy system using the original six input variables. FM is the neurofuzzy system with factor analysis using the maximum likelihood estimation as a method of the extraction with the newly extracted components or factors. FMV is the neurofuzzy system with factor analysis applying varimax rotation and maximum likelihood estimation as a method of the extraction. FC is the neurofuzzy system with the factor analysis using the principal components as a method of the extraction. FCV is the neurofuzzy system with factor analysis using varimax rotation and principal components as a method of the extraction. PCR is the neurofuzzy system with principal component analysis using the correlation. Finally, PCV is the neurofuzzy system with principal component analysis using the covariance. The reduced components are also examined with the six original measurements types using the statistical categories such as root means square (RMS), standard deviations (STD), mean of absolute distance (MAD), statistical index (EWI) and error rate (ERR).

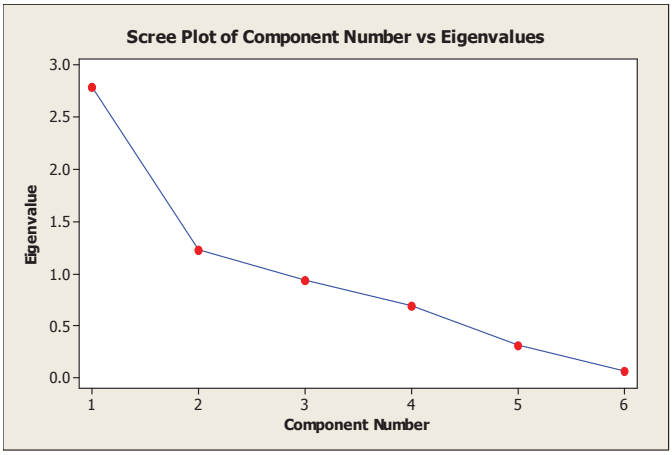


Fig. 7 The relationship between Components and Eigenvalues

TABLE 1 Statistical Analysis between extracted components with estimated values and the original values using neurofuzzy systems

	TRMS	STD	MAD	EWI	ERR
ORG	0.4844	1.9149	0.4791	2.8794	0.2918
FC	0.2982	0.6190	0.2949	1.2130	0.2575
FCV	0.6765	3.6542	0.6691	5.0008	0.3564
FM	0.1220	0.7777	0.1206	1.0204	0.0601
FMV	0.6974	3.6833	0.6897	5.0714	0.4219
PCR	0.3484	0.8272	0.3446	1.5212	0.2815
PCV	0.2799	0.5994	0.2769	1.1571	0.2433

From TABLE 1, only five newly extracted components are applied to predict the plasma beta-carotene levels using five inputs neurofuzzy system. For the root mean square (TRMS), FM evaluates the best matches. In the standard deviation (STD), PCV's evaluation draws the best performance. In the category of mean of absolute distance (MAD), FM evaluates the closest prediction of the plasma beta-carotene levels. In overall, using the category of equally weight index (EWI), FM shows the best result among the other techniques. Finally, for the error rate (ERR), FM shows the best result among other neurofuzzy systems.

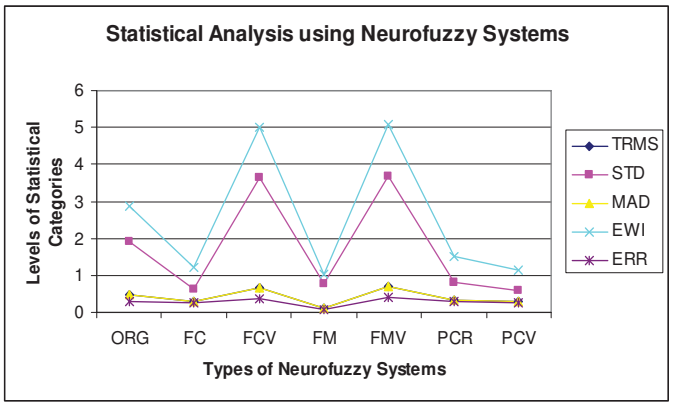


Fig. 8 Comparison of Statistical Analysis using extracted components through neurofuzzy systems

Fig. 8 plots the statistical evaluation using five different categories with the comparison based upon the suggested statistical measurements. RMS, STD, MAD, EWI, and ERR in Fig. 8 stand for the statistically evaluation with the six original components and five newly extracted components using neurofuzzy systems, respectively.

The following statistical categories evaluate the prediction of the performance using the neuro fuzzy systems with reduced data models:

Total Root Mean Square (TRMS): Total Root Mean Square for the distance between the original output and the estimated output using the same testing data through the neurofuzzy system.

$$RMS = \frac{\sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}}{n-1} \tag{7}$$

where x_i is the estimated value and y_i is the original output value. From TABLE 1, FM shows the best prediction using the neurofuzzy system generated by the reduced factors.

Standard Deviation (STD): The standard deviation presents the differences between the original output and the predicted output using the same testing data. From TABLE 1, PCV shows the best prediction of the plasma beta-carotene levels through the neurofuzzy system generated by the newly reduced components.

Mean of the Absolute Distances (MAD): The mean of the absolute distances measures the difference between the original output and the predicted output using the absolute difference values with the same testing data. From TABLE 1,

FM shows the best prediction of the plasma beta-carotene levels through the neurofuzzy system generated by newly extracted factors.

Equally Weighted Index (EWI) [10]: The index value from the summation of the values with multiplying the statistical estimation value is calculated by its equally weighted potential value for each statistical category field. The value, which is close to 0, is the better results. From TABLE 1, PCV shows the best prediction of the plasma beta-carotene levels through the neurofuzzy system generated by the newly extracted components.

Error Rate (ERR): the error rate between the predicted plasma beta-carotene levels and the original plasma beta-carotene levels is calculated. From TABLE 1, FM shows the best prediction of the plasma beta-carotene levels through the neurofuzzy systems generated by the newly extracted factors.

6 Conclusions

The presented paper describes how efficiently to determine the level of the plasma concentration using the estimation of the beta carotene level from the original data as well as the various reduced dimensionality data without losing any significant meaning through evaluating the system output with the various neurofuzzy systems. The plasma concentration data are employed and identified by using five statistical measurements. The evaluation values from the five statistical categories are relatively lower than the expected level because the sampled data which used for the various statistical techniques to diminish the dimensionality were relatively uncorrelated among the adapted measurement types. In overall, the dimensional scaling with applying the factor analysis with maximum likelihood estimation shows the best performance in the equally weighted index and the error rate through the evaluation using the neurofuzzy system.

Acknowledgment

The original data is from the website: http://lib.stat.cmu.edu/datasets/Plasma_Retinol and a related reference is from the journal paper written by Nierenberg and et al. [8].

7 References

[1] Kristin L. Sainani, "Introduction to principal components analysis," Physical Medicine and Rehabilitation (PM&R), Vol. 6 Issue 3, pp. 275-278, March, 2014.

[2] Timothy A. Brown, Confirmatory Factor Analysis for Applied Research, Second Edition, The Guilford Press, New York, NY, 2015.

[3] I. Myung, "Tutorial on maximum likelihood estimation," Journal of Mathematical Psychology, 47, pp. 90–100, 2003.

[4] Greene-2140242 book, New York University, (Chapter 14), 2010. http://pages.stern.nyu.edu/~wgreene/DiscreteChoice/Readings/Greene-Chapter-16.pdf

[5] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," Psychometrika 23, pp. 187–200, 1989.

[6] C. Lin, and C. Lee, Neural Fuzzy Systems, Upper Saddle River: Prentice-Hall, (Part II and III), 1996.

[7] J. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," IEEE Trans. Systems, Man & Cybernetics, Vol. 23, pp. 665−685, 1993.

[8] D. Nierenberg, T. Stukel, J. Baron, B. Dain, and E. Greenberg, Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology, Vol. 130, pp. 511-521, 1989. http://lib.stat.cmu.edu/datasets/Plasma_Retinol

[9] N. Cliff, "The Eigenvalues-Greater-Than-One Rule and the Reliability of Components," Psychological Bulletin, Vol. 103, No. 2, pp. 276-279, 1988.

[10] D. Nam, and H. Singh, "Material processing for ADI data using multivariate analysis with neuro fuzzy systems," Proceedings of the ISCA 19th International Conference on Computer Applications in Industry and Engineering, Las Vegas, Nevada, pp. 151–156, Nov. 2006.

[11] I. Jolliffe, Principal Component Analysis (2nd ed.), New York: Springer, (Chapter 2 and 3), 2002.

[12] R. Rummel, Understanding Factor Analysis, Journal of Conflict Resolution, Vol. 11, No. 4, pp. 444-480, 1967.

[13] The Pennsylvania State University Tutorial for Varimax Rotation, 2004, http://sites.stat.psu.edu/~ajw13/stat505/fa06/17_factor/13_factor_varimax.html

[14] K. Ramesh, A. P. Kesarkar, J. Bhate, M. Venkat Ratnam, and A. Jayaraman, "Adaptive neuro-fuzzy inference system for temperature and humidity profile retrieval from microwave radiometer observations," Atmospheric Measurement Techniques, Vol. 8, pp. 369–384, 2015.

[15] Konstantin Kashin, "Statistical Inference: Maximum Likelihood Estimation," Notes from Spring 2014, http://www.konstantinkashin.com/notes/stat/Maximum_Likelihood_Estimation.pdf

A Scalable Strategy for Mining Association Rules under Grids

R. Tlili[1], K. ElBedoui[2], and Y. Slimani[3]
[1]Department of Computer Science, Faculty of Sciences of Tunis (FST), Campus Universitaire, Tunis, Tunisia
[2]Department of Computer Science, Faculty of Sciences of Tunis (FST), Campus Universitaire, Tunis, Tunisia
[3]Department of Computer Science, Higher Institute of Multimedia Arts of Manouba (ISAMM), Campus Universitaire, Manouba, Tunisia

Abstract— *Sequential Association Rule Mining (ARM) algorithms are characterized by a high computational complexity due to two facts: (i) they have to mine a very large search space (ii)they have high demands of database access. Association rule mining technique have progressively been adapted to large-scale systems in order to benefit from the large-scale computing capabilities and the huge storage capacity provided by theses systems. Performance issues (i.e, efficiency and scalability) are determinant factors for association rule mining algorithms [1].*

In this paper we present an important part of our multilevel strategy that aims to improve the scalability of distributed ARM algorithms. Our main goal is to obtain a running time that grow linearly in proportion with the size of the database, given the available system resources (i.e., available computing nodes, their main memory and their disk space, etc.). The French research grid "Grid'5000" is used as our experimental test-bed.

Keywords: Distributed association rule mining, running time, Scalability, Grid'5000.

1. Introduction

Association rules mining is one of the most well studied data mining techniques [2]. It was first introduced by Agrawal for transaction data analysis [3]. Today, this technique is used in a wide variety of applications such as intrusion detection, heterogeneous genome data, mining remotely sensed images/data, product assortment decisions, telecommunication networks, market and risk management, etc [4].

Extracting useful knowledge from data sets measuring in Petabytes is a challenging research area for the data mining community. Sequential approaches suffer from a performance problem due to the fact that they have to mine large and also in the majority of real-world cases distributed databases [5], [6], [7]. Parallelism is introduced as an important solution that could improve the response time and the scalability of these approaches. However, because the parallelization process is not a trivial straightforward process, it introduces a plethora of new problems including the scalability problem [8], [9].

Although Grid computing systems share many aspects with parallel and distributed approaches, there are platform peculiarities and requirements implying extra efforts and new methodologies to deal with the heterogeneity of such systems. Running classical parallel and distributed algorithms under Grid systems will degrades their performance due to the load imbalance that appears between resources during execution time [9].

The research work of our paper is targeted to cope with the challenges brought by running association rule mining algorithms under grid computing environments. We embedded our multilevel scalability improving strategy into different parallel association rule mining algorithms. The rest of the paper is organized as follows: Section 2 introduces The multilevel strategy for improving the ARM algorithm scalability. Section 3 presents the processing level of our strategy. Sections 4, 5 and 6 gives the details and the different stages of the processing level. Experimental results obtained from implementing this strategy are showed in section 7. The paper concludes with section 8.

2. The multilevel strategy for improving the ARM algorithm scalability

The objective of the researcher in the domain of knowledge discovery is to design high performance ARM algorithms. Efficiency and scalability are the two determinant factors for the performance of ARM algorithms. For an algorithm to be scalable, its running time should grow linearly in proportion to the size of the database, given the available computing system resources [2]. So, scalability has two main dimensions: (i) The very large and incremental size of the databases that have to mined. An efficient ARM algorithm must be capable of extracting knowledge in a reasonable time even when the size of the database is continuously increasing. (ii) The ARM algorithm must scale very well when the number of computing resources provided by the grid increases. It must be capable of dealing with different synchronization barriers and results consolidations that must be done by the end of each iteration, without degrading the overall execution time. Both the dynamic behavior of ARM algorithms and grid characteristics have an impact on the design of our multilevel scalability improving strategy.

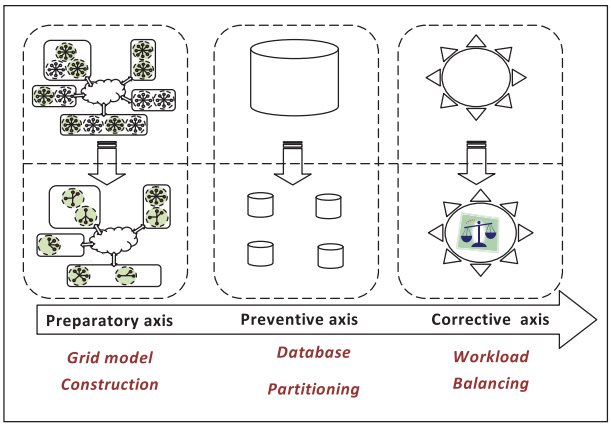


Fig. 1: Workload balancing strategy axis

The scalable strategy is mainly based on three levels or axes of interference. Figure 1 depicts the sequence formed by these levels:

1) **Preparatory level**: Responsible of adapting the distributed ARM algorithm to the hierarchical grid model [10].
2) **Preventive level**: Since the workload imbalance has a direct impact on the scalability of the distributed algorithm [1], [11], the intervention on this axis aims to minimize the probability of workload imbalance during processing time. This is achieved by introducing a new data partitioning approach which takes into account the particularities of both ARM technique and grid environment [12].
3) **Corrective level**: This axis represents the processing phase of the multilevel scalability improving strategy. In this level the workload imbalance observed during the execution is corrected dynamically. This level is very critical and delicate, because if workload imbalance corrections are not delicately measured they could increase the overall execution time instead of decreasing it. These corrections are the key idea that help the distributed ARM algorithm to scale with the increasing size of the database and also the increasing number of computing nodes incorporated in execution. This level will be explained in details in what follows.

These three levels form together a complementary homogeneous sequence that aims to ensure a proper execution of the ARM algorithm under a grid computing environment, while maintaining workload balancing and providing rational exploitation of all available computing resources. The third level (i.e., the corrective level also called the processing level) is the longest phase in the process with a dynamic and evolving behavior. The processing level will be explained in details in what follows.

3. Processing level

The specific characteristics of ARM algorithm with those of the computing environment (Grid) must be taken into account. While association rule mining method is based on global supports, we are only disposed by partial supports at the end of each iteration. This is due to the fact of dataset distribution. Synchronization barriers are then necessary in order to obtain global supports of itemsets. The difference in processing capabilities of the components of the grid joined with the difference in the amount of work needed for each data partition lead to high time delays behind synchronization barriers. Our goal is to reduce the idle time periods of different computing resources by inducing workload balancing. This will allow the distributed algorithm to benefit as much as possible from the available computing power. The processing phase of our strategy will be presented in three parts:

1) What to balance: defining the concept of "workload" in ARM algorithms.
2) Where to balance: defining the levels of workload balancing (i.e., load balancing hierarchies).
3) How to balance: the modules of the workload balancing application.

4. What to balance ?

Our goal is to control the workload of the ARM algorithm running under the grid and to balance it if necessary. So, it is necessary to start by defining the work (i.e., ARM tasks) that we want to control and balance and then finding a way to measure it. Each entity of our grid (node, cluster and site) has a specific workload that changes (increases or decreases) during the processing phase. As the node is the basic entity of a grid, we first define the workload at a given node.

4.1 Node level

The workload of a node is the ratio between the computing capacity of the node (expressed in term of processing units also called intructions per second) and the amount of work to be done (expressed in the number of transactions to be processed and the number of generated candidate itemsets). The triplet of information (capacity calculation, number of transactions, number of candidate itemsets) allows us to deduce the necessary elements to define the workload of a node. The most important criterion for a node is not the amount of work already done but rather the amount of remaining work. Hence, we define the workload of a node, denoted by WL_{ijk}, as the amount of work that still have to be done, expressed in number of processing units needed to perform the remaining work. Starting from the speed of a node Nd_{ijk}, expressed in instructions per second, and knowing the number of processed and remaining transactions of this node at a given instant, and the time interval separating two periods of workload information

calculation, we define first $PuTr_{ijk}$ which is the number of processing units required to process a transaction Tr_i.

$$PuTr_{ijk} = IS(Nd_{ijk})/NPtr_{ijk} \quad (1)$$

Thus, we can deduce the workload of a node as follows:

$$WL_{ijk} = PuTr_{ijk} * NRT_{ijk} \quad (2)$$

To ensure a permanent control over all entities of the grid, each node send periodically to the coordinator of the cluster to which it belongs, its workload WL_{ijk}. This communication is done under the form of a set of information, which we call "Node Workload State Vector" ($NWSV_{ijk}$).

4.2 Cluster level

For a cluster, two important information are calculated:

- The workload of the cluster: as each cluster Cl_{ij} is composed of a fixed number of nodes. Then, its workload can be deducted from summing the workloads of its nodes received through the $NWSV_{ijk}$ of each Nd_{ijk}.
- The average workload of the cluster is defined by the following equation:

$$WL_{ij} = \sum_{k=1}^{N_{ij}} WL_{ijk}/N_{ij} \quad (3)$$

We define the computational workload of the cluster as the average workload of its nodes instead of the sum of these workloads because different clusters do not have the same number of nodes and hence the sum of the workloads can not be an eligible factor of comparison. So, the average workload allows us to deduce the workload state of the cluster regardless of the number of nodes constituting it.

At this level it is important for the coordinator $Coord(Cl_{ij})$ of the cluster Cl_{ij} to have an idea about how much the workloads of the nodes under its control are spread out with respect to the computational workload WL_{ij} of the cluster. For this, we define an interval based on the standard deviation.

The standard deviation of a cluster Cl_{ij}, denoted by σ_{ij}, shows how much variation or "dispersion" exists from the average workload. The standard deviation is the square root of the variance. The variance is a measure of how far the set of workloads are spread out and is calculated as the average of the squared differences from the mean.

Therefore, the workload of the cluster and the standard deviation calculated from the workloads of different nodes, allow the coordinator of the cluster to define two important thresholds as follows:

- The standard deviation associated to cluster Cl_{ij} is defined by:

$$\sigma_{ij} = \sqrt{1/N_{ij} \cdot \sum_{k=1}^{N_{ij}} (WL_{ij} - WL_{ijk})^2} \quad (4)$$

- The $ThresMax$ determines the threshold above which a node is considered to be overloaded and needs a workload balancing phase.

$$ThresMax = WL_{ij} + \sigma_{ij} \quad (5)$$

- The $ThresMin$ allows to determine if a node is underloaded or not. If it is the case, it can receive an amount of work from an overloaded node.

$$ThresMin = WL_{ij} - \sigma_{ij} \quad (6)$$

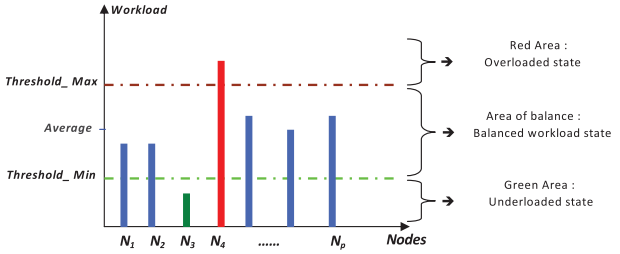


Fig. 2: Workload states of computing nodes

These three values allow to the coordinator of a cluster Cl_{ij} to divide its nodes into three classes (see Figure 2):

- Underloaded class formed by all nodes with a workload below $ThresMin$ (like the node $N3$ in Figure 2). The nodes of this class have the capacity of receiving additional work and therefore they represent the "Green Area" for the coordinator of the cluster.
- Balanced class: groups nodes with a workload between $ThresMin$ and $ThresMax$ (like nodes $N1$ and $N2$ in Figure 2). The workload of these nodes is close to the average workload of their cluster; therefore they are in a balance state.
- Overloaded class which is formed by all nodes having a workload above the $ThresMax$ (like node $N4$ in Figure 2). The workload of these nodes need to decrease their workloads through workload balancing operations. They represent the "Red Area" for the coordinator of the cluster. During a workload balancing operation, these nodes send part of their work to underloaded nodes.

4.3 Site level

The same logic applied to the cluster level is also used to define the load of a given site. Therefore, the workload of a site is equal to the average load of

The workload of site S_i will be given by the following equation:

$$WL_i = \sum_{j=1}^{M_i} WL_{ij}/M_i \quad (7)$$

Hence, the same principle used in classifying the nodes of a cluster into underloaded, overloaded and balanced will be applied to classify the clusters of a site as follows:

- The standard deviation associated to site S_i is defined by:

$$\sigma_i = \sqrt{1/M_i \cdot \sum_{j=1}^{M_i} (WL_i - WL_{ij})^2} \quad (8)$$

- Clusters with workloads greater than $ThresMax$ are considered to be overloaded.

$$ThresMax = WL_i + \sigma_i \quad (9)$$

- Clusters with workloads less than $ThresMin$ are considered to be underloaded.

$$ThresMin = WL_i - \sigma_i \quad (10)$$

- Clusters between $ThresMin$ and $ThresMax$ are in a balanced state.

4.4 Grid level

At this level,the workload of a grid G is considered to be the average workload of its sites and is calculated by the following equation:

$$WL_G = \sum_{i=1}^{T} WL_i/T \quad (11)$$

Where, T is the total number of sites in the grid G and WL_G is the computational workload of G.

The sites of the grid G could be classified to underloaded, overloaded and balanced as follows:

- The standard deviation associated to the grid G_i is defined by:

$$\sigma_G = \sqrt{1/T \cdot \sum_{i=1}^{T} (WL_G - WL_i)^2} \quad (12)$$

- Sites with workloads greater than $ThresMax$ are considered to be overloaded.

$$ThresMax = WL_G + \sigma_G \quad (13)$$

- Sites with workloads less than $ThresMin$ are considered to be underloaded.

$$ThresMin = WL_G - \sigma_G \quad (14)$$

- Sites between $ThresMin$ and $ThresMax$ are in a balanced state.

5. Where to balance: defining the workload balancing hierarchy ?

Based on the hierarchical modeling of the grid, defined in the preparatory level of our scalability improving strategy [10], is build a hierarchical and distributed load balancing plan. This plan is constituted of the following three levels:

1) **Intra-cluster workload balancing:** this first level concerns the migration of work between nodes of the same cluster. The coordinator of the cluster $Coord(Cl_{ij})$ privileges primarily this local workload operation. It seeks through workload state vectors of its nodes to find the appropriate underloaded nodes that can alleviate the load of an overloaded node. If workload imbalance still persists, the coordinator of the cluster moves to the next level.
2) **Intra-site workload balancing:** this level depends on the migration of work between clusters within the same site. The site coordinator decides to initiate a workload balancing operation based on load information (workload state vectors) sent periodically by the coordinators of clusters. The site coordinator gives priority for balancing the workload by redistributing it locally between clusters which are under its control. This approach of locality aims to reduce the communication cost, by avoiding inter-sites communications which use the WAN network.
3) **Inter-sites workload balancing:** the workload balancing, at this third level, is triggered when some sites coordinators fail in their attempts to balance workload locally through their respective sites. The failure of local workload balancing may be due to the saturation of the site, or to insufficient charge offer induced by the lightly-loaded cluster with respect to the request formulated by overloaded nodes. In this case, the site coordinator tries to find another site which is able of accepting the current overload. This search is accomplished by negotiating the transfer of candidate itemsets, transactions or both from the overloaded site to the underloaded one.

A negotiation process is launched to find the appropriate nodes (either local or remote node) able to relieve the overloaded nodes. This process takes into account the communication factor in order to guaranty that the workload balancing process will improve the performance of the ARM algorithm running under the grid environment.

6. How to balance: application modules

The role of a node is limited to two basic functions:

- A support counting function which is ensured by the "DDMA" module.
- A monitoring function which consists of periodically generating the workload state vector of the node and sending it to the cluster coordinator. Other communication aspects between a computing node and its coordinator of cluster may be useful as receiving new candidate itemsets or a reduction in the number of candidate itemsets during execution. All these functionalities of monitoring and communication are provided by the module "Monitoring".

The functionalities of a Coordinator (cluster/site) are more complex than the ones at the node level:

- An initialization function, ensured by the "Initialization" module. This module covers the preprocessing phase of our workload balancing approach.
- A consolidation function where partial results provided by the "DDMA" module are exchanged. This function is done by the "DDMA" module.
- A monitoring function on the entities which are under the responsibility of the relevant coordinator: this function is provided by the module "Monitoring". In addition to that, this module ensures all internal and external communications of the coordinator with other grid entities, decision making when local workload balancing is needed or when receiving an external demand of inter-sites workload balancing.
- A workload balancing function which is performed by "Load Balancing" module. This module executes the workload balancing decisions made by the "Monitoring" module.

The previously defined structures have a well-defined logic of communication during execution time. The modules of each structure collaborate internally (within the same entity) and externally (with other entities of the grid) via communication links as shown in Figure 3.

All communications are based on a message exchange system that has been established for the purpose.

The module "Initialization" initializes the launching of execution by preparing the partitioning and the distribution of the transactional database. This module communicates with its corresponding "Initialization" module of the entity which is directly related to it: Site-Cluster and Cluster-Node.

The module "DDMA" executes the various tasks of the ARM algorithm like the generation of candidate itemsets, computing their frequencies, exchanging and consolidation of results between different coordinators. This module communicates with its corresponding "DDMA" module of other entities and also with the module "Workload Balancing" in the same entity to accept or delegate a new part of data to be processed in case of workload imbalance.

The module "Monitoring" controls the progress of execution at each grid entity. It ensures the exchange and the control of workload state vectors exchanged during execution. This module communicates with two modules: "DDMA" and "Workload Balancing" of the same entity.

The module "WorkLoad Balancing" intervenes when the module "Monitoring" detects a workload imbalance. This module involves different workload balancing levels provided by our strategy: intra-cluster, intra-site and inter-sites workload balancing. This module communicates with the module "DDMA" of its entity in case of transactions or candidates migration (i.e., addition/deletion of data or candidates to/from the entity). It also communicates with the module "Load Balancing" of other entities when remote workload balancing is needed to negotiate migration.

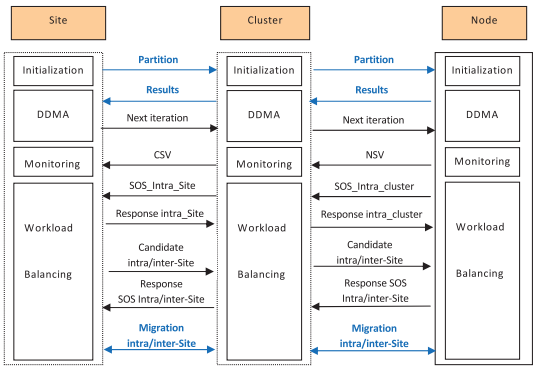


Fig. 3: Flow of communication between different modules of the workload balancing strategy

Figure 3, illustrates an example of exchanged messages between the entities of the grid. In this flow we distinguish two kind of messages:

- Data messages:
  - Partition: partitions of the transactional database.
  - Results: frequencies of itemsets.
  - Migration intra/inter-Site: candidate itemsets migrating between nodes of the same cluster and transactions migrating between clusters/sites.
- Control messages:
  - CSV: Cluster workload State Vector.
  - NSV: Node workload State Vector.
  - SOS_Intra_Site: Intra-site workload balancing request. This type of message is initiated by a site coordinator, to all entities under its responsibility, to search primarily for a possibility of local balancing. This is an internal message of the site.
  - SOS_Inter_Site: Inter-sites workload balancing request. This message is launched by a site coordinator for other site coordinators, to search for the possibility of balancing inter-sites. This message is triggered when the message "Response SOS Intra-

Site" returns a failure, which means that there is no possibility of local workload balancing.

In the next section, we will evaluate the performance of the perviously detailed strategy using the Grid'5000 platform.

7. Performance evaluation

The performance evaluations presented in this section were conducted on Grid'5000 [13], a dedicated reconfigurable and controllable experimental platform featuring 13 clusters, each with 58 to 342 PCs, interconnected through Renater (the French Educational and Research wide area Network). It gathers roughly 5000 CPU cores featuring four architectures (Itanium, Xeon, G5 and Opteron) distributed into 13 clusters over 9 cities in France (Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Sophia-Antipolis, and Toulouse) [13].

In order to do our tests, we implemented a parallel version of the DCI algorithm [7]. According to tests conducted during the FIMI workshop [1], DCI algorithm is found to be one of the fastest ARM algorithms. It can be used efficiently to find frequent itemsets even with very low support values. This algorithm adopts a hybrid approach to determine the supports of frequent itemsets, by exploiting a counting-based method during first iterations (i.e., it scans a horizontal transactional database layout), and an intersection-based method when the dataset can fit into main memory (i.e., during this phase, it scans a vertical transactional database layout). This algorithm also uses simple data structures with direct access for storing candidate itemsets. Then, we embedded our multilevel strategy within the parallel DCI algorithm. The datasets used in tests are synthetic, and are generated using the IBM-generator. Table 1 shows the datasets characteristics.

Table 1: Transactional database characteristics

Database	Characteristics		
	# Items	Avg. Trans. Length	# Transactions
DB5250M12235T	250 000	250	12 235 857

The following grid configurations are used:

- For the case of 2 and 4 nodes, there are: 1 coordinator, 1 site, 1 cluster.
- For the case of 8 nodes, there are: 2 coordinators, 1 site, 2 clusters, 10 cores.
- For the case of 16 nodes, there are: 4 coordinators, 1 site, 4 clusters, 20 cores.
- For the case of 32 nodes, there are: 7 coordinators, 2 sites, 5 clusters, 39 cores.
- For the case of 64 nodes, there are: 12 coordinators, 3 sites, 9 clusters, 76 cores.

[1] http://fimi.cs.helsinki.fi/

The table displayed in figure 4 illustrates the execution time, the speedup and the efficiency of the DCI algorithm with and without the multilevel scalability improving strategy. For abbreviation, "NoLB-DCI" refers to the parallel version of DCI algorithm without workload balancing, while "LB-DCI" refers to the workload balanced DCI. From figure 5, we can notice that the execution time, of NoLB-DCI with 64 nodes is equal to 3480 sec, while the execution time of LB-Apriori is equal to 2630 sec. This gives us a gain in the time of execution of about 24.43%. Also, as the number of computing resources increases from 2 to 64 the execution time of LB-DCI is remarkably decreasing. The optimal execution time of NoLB-DCI, for DB5250M12235T dataset, is obtained with 64 nodes. Beyond this number of nodes, the execution time starts increasing again. With LB-DCI, we still have an increasing improvement in execution time. This shows that LB-DCI scales very well with the increase in the number of computing nodes. As depicted in figure 6, the speedup of NoLB-DCI with 64 nodes is equal to 48.01, while the speedup of LB-DCI is equal to 63.53. This means that the LB-Apriori benefits more from available computing resources. The number of computing nodes incorporated in execution must be proportional to the amount of work needed by a specific database and with a specific support value. This means that, when an optimal value of execution time is reached experimentally, then there is no need to further increment the number of used nodes. The need of workload balancing increases as the number of nodes increases. From figures 5, 6 and 7 we can notice that our multilevel strategy offers great savings in processing time and improves the scalability of the parallel ARM algorithm.

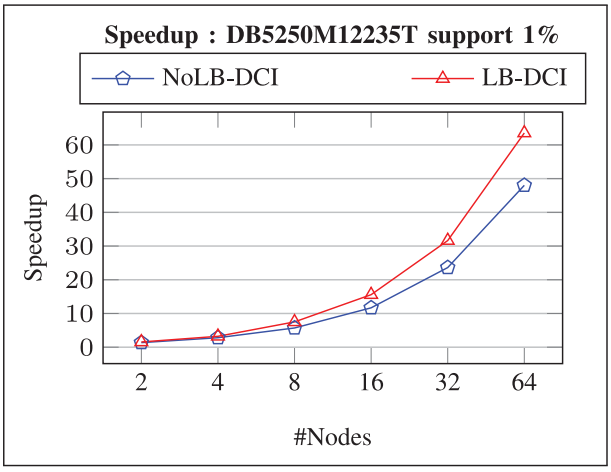


Fig. 6: Speedup of DCI algorithm with and without load balancing

DB5250M12235T support 1%	# Nodes	1	2	4	8	16	32	64
	# Cores		3	5	10	20	39	76
	NoLB-DCI	167089	123739	59402	29339	14303	7073	3480
	LB-DCI	167089	107467	45463	22278	10724	5283	2630
	Gain %	00.00	13.15	23.47	24.07	25.02	25.31	24.43
	speedup NoLB-DCI		01.35	02.81	05.70	11.68	23.62	48.01
	speedup LB-DCI		01.55	03.68	07.50	15.58	31.63	63.53
	Effeciency NoLB-DCI		00.68	00.70	00.71	00.73	00.74	00.75
	Effeciency LB-DCI		00.52	00.74	00.75	00.78	00.81	00.84

Fig. 4: Execution time, speedup and efficiency for DB5250M12235T

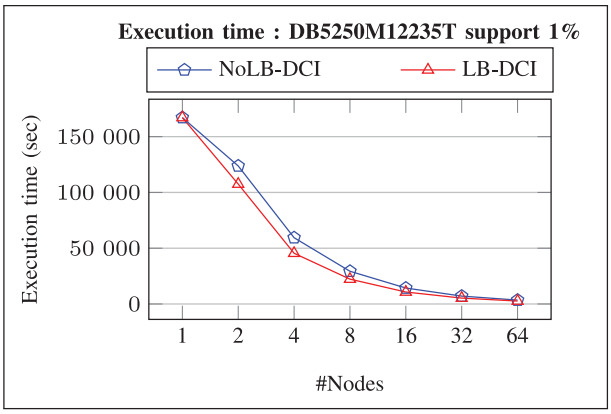


Fig. 5: Execution time of DCI algorithm with and without load balancing

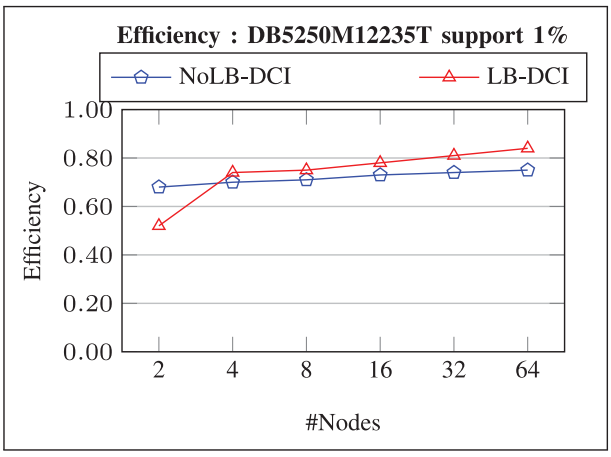


Fig. 7: Efficiency of DCI algorithm with and without load balancing

8. Conclusions

ARM algorithms are both data and computationally intensive. The expected benefits from embedding association rule mining into a grid environment are execution time acceleration and scalability. With workload imbalance, the distributed ARM algorithm can only execute at the speed of the most heavily loaded computing node. In this paper we proposed a multilevel strategy that corrects the skew in workload that occurs during the execution of the ARM algorithm under the goal of improving the scalability. Experiments, using Grid'5000 platform, showed that our approach can successfully maintain the workload balance and this leads to an improvement in the scalability in term of the amount of treated data. It also helps the distributed algorithm in benefiting as much as possible from the large-scale computing capabilities provided by a grid environment (i.e., the scalability in terms of computing resources).

References

[1] M.J. Zaki. Parallel and distributed association mining : A survey. *Concur-rency, spécial issue on Parallel Mechanisms for Data Mining*, 7(4):14–25, December 1999.

[2] J. Han and M. Kamber. *Data mining: concepts and techniques.* Maurgan Kaufman Publishers, 2000.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining associations rules in large databases. In *the 20th Int. Conf. on Very Large Data Bases*, pages 478–499, September 1994.

[4] S. Kotsiantis and D. Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

[5] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, May 1993.

[6] T. Shintani and M. Kitsuregawa. Hash-based parallel algorithms for mining association rules. In *Proceedings of the International Conference on Parallel and Distributed Information Systems*, pages 19–30, CA, USA, 1996.

[7] S. Orlando, P. Palmerini, R. Perego, and F. Silvestri. A scalable multi-strategy algorithm for counting frequent sets. In *Proceedings of the 5th Workshop on High Performance Data Mining*, Washington, USA, April 2002.

[8] J. Chattratichat, J. Darlington, M. Ghanem, H. Hüning Y. Guo, M. Köhler, J. Sutiwaraphun, H. W. To, and D. Yang. Large scale data mining: the challenges and the solutions. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 1997.

[9] Y. B. Ma, S. H. Song T. Y. Kim, and J. S. Lee. Analysis and experimentation of grid-based data mining with dynamic load balancing. In *Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA'09)*, pages 561–568, 2009.

[10] R. Tlili and Y. Slimani. A hierarchical dynamic load balancing strategy for distributed data mining. *International Journal of Advanced Science and Technology*, February 2012.

[11] V. Fiolet and B. Toursel. Distributed data mining. in scalable computing : Practice and experiences (scpe). *Internet-based Computing*, 6(1):99–10, March 2005.

[12] R. Tlili and Y. Slimani. A novel data partitioning approach for association rule mining on grids. *International Journal of Grid and Distributed Computing*, December 2012.

[13] F. Cappello, E. Caron, M. Dayde, F. Desprez, Y. Jegou, P. Vicat-Blanc Primet, E. Jeannot, S. Lanteriand J. Leduc, N. Melab, G. Mornet, B. Quetier, and O. Richard. Grid'5000: a large scale and highly reconfigurable grid experimental testbed. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, pages 99–106, Washington, USA, November 2005.

SESSION

FEATURE EXTRACTION AND RELATED ALGORITHMS + INFORMATION RETRIEVAL and ONTOLOGY BASED METHODS

Chair(s)

TBA

ODISEES: Ontology-driven Interactive Search Environment for Earth Sciences

Matthew T. Rutherford[1], **Elisabeth B. Huffer**[2], **John M. Kusterer**[3], **Brandi M. Quam**[4]
[1,2] Lingua Logica LLC, NASA Langley, Atmospheric Science Data Center, Hampton, VA, United States
[3,4] Science Directorate, NASA Langley, Atmospheric Science Data Center, Hampton, VA, United States

Abstract— *This paper discusses the Ontology-driven Interactive Search Environment for Earth Sciences (ODISEES) project currently being developed to aid researchers attempting to find usable data among an overabundance of closely related data. ODISEES' ontological structure relies on a modular, adaptable concept modeling approach, which allows the domain to be modeled more or less as it is without much concern for terminology or external requirements. In the model, variables are individually assigned semantic content based on the characteristics of the measurements they represent, allowing intuitive discovery and comparison of data without requiring the user to sift through large numbers of data sets and variables to find the desired information.*

Keywords: ontology, semantics, data discovery, search

1. Introduction

Over the past few decades, researchers have often been faced with a unique and very modern problem: the overabundance of usable, relevant data. The exponentially shrinking cost and size of computer hardware components has enabled the storage of vast quantities of data, and the emergence of the Internet has enabled near-instantaneous dissemination of this data. As a result, finding precisely the right data can be like searching for the proverbial needle in a haystack. Oftentimes, researchers must sift through large volumes of closely related information in order to uncover the desired usable information buried there. This is, in many ways, the opposite of what researchers have had to deal with in the past where data was generally much less abundant and more difficult to assess. As such, the scientific and information technology communities face the ever-growing challenge of storing, managing and distributing vast amounts of data and providing user-friendly tools to the entities that hope to use the information therein. The Ontology-Driven Interactive Search Environment for Earth Sciences (ODISEES) project seeks to offer an alternative to traditional methods of data discovery, organization, and access.

There is a wealth of Earth science data—atmospheric, geological, meteorological, hydrological, etc.—that has grown rapdily over the past few decades. These data are produced through a variety of collection methods and technologies, and are interpreted by scientists and researchers from a wide, heterogeneous set of disciplines for a similarly large and varied set of purposes. Earth science researchers are often faced with two significant, recurring challenges:

1) Finding data products that are immediately relevant to their research
2) Quickly noticing and understanding the similarities and differences among closely related products and assessing their suitability for a particular purpose

As one of the Distributed Active Archive Centers (DAACs) under the umbrella of NASA's Earth Observing System Data and Information System (EOSDIS), the Atmospheric Science Data Center (ASDC) at the NASA Langley Research Center is a steward of large amounts of Earth science data. The ASDC's purpose [1] is to make its petabytes of data holdings easily available to the public, serving a variety of users, including scientists and researchers as well as the general public. Given that scientists and laymen alike are actively accessing and using its archives, the ASDC is met with the unique challenge of catering to a user base with inconsistent knowledge of its data holdings.

The use of ontologies and semantics has emerged as one of many solutions to address these issues. Our implementation of this solution, ODISEES, is being developed to provide researchers with tools for discovering and assessing available ASDC data products.

2. Ontology

There are many interpretations of the term "ontology". Long used by philosophers as a conceptual tool for studying the conditions of existence and for classifying that which exists, ontology has more recently been interpreted and implemented by computer and information scientists—primarily as a computer-readable artifact that can represent the entities and relationships in a domain of interest.

Historically, ontology-based solutions have often started out ambitiously. Researchers in Artificial Intelligence, for instance, once looked to ontology and deductive reasoning systems in an attempt to create truly intelligent systems that could overcome the limitations of expert systems. The Cyc project [2] [3] is one such project, designed to represent the contextual common sense knowledge that humans take for granted when they engage in deductive reasoning. However, backlash over the years against AI in general, deductive reasoning systems like Cyc in particular [4] [5] [6], and top

down ontologies, which try to impose context-independent structure, resulted in less ambitious use cases for ontology. At the same time, the Semantic Web was developing and ontologies emerged as a means for controlling vocabularies and bolstering the exchange of information on the Internet [7]. The Resource Description Framework (RDF) [7] [8], which effectively merged ontology and XML, was developed. Since then, the word "ontology" has largely been used to refer to controlled vocabularies that are used to provide semantic content for marking up data objects for the Internet. In this context, the reasoning and inference capabilities that were previously critical to AI applications became less important.

While less ambitious uses of ontology have become more widely accepted, the term "ontology" continues to be subject to multiple interpretations. One of the more popular definitions of ontology in an information and computer science context comes from Tom Gruber [9]: "...an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)." Some liberal interpretations can include taxonomies, relational models, and XML, but these do not provide any semantics. At the other end of this interpretive spectrum are full-fledged logical theories of domain that rely on first-order, second-order, and even modal logic. In developing ODISEES, we were interested in the latter, more robust interpretation and application of ontology: artifacts that model a domain with a good deal of precision and leverage the inferential powers of first and second-order logic.

3. Methodology

ODISEES relies on the ontological classification of measured phenomena represented in ASDC data products, using the resulting characterization of the data to enable effective, expedited discovery and comparison of said data. It was designed to perform three primary tasks in service to data search and discovery:

1) Model the set of objects and concepts that make up the Earth science domain and the relationships among them in order to provide a common domain model to impart meaning to the terms used to describe the domain
2) Identify and define the terms used by specialized user groups within the Earth science community, mapping these terms to the common domain model and creating computer-readable definitions of community-based terms
3) Be usable by humans, databases, and applications that need to interpret model to impart meaning to the terms used to describe the domain

The ODISEES search system is comprised of open-source, commercial-off-the-shelf (COTS), and custom software components that interact with an Earth science ontology and metadata repository to offer long-term solutions to the problem of data discovery in the era of Big Data. The ODISEES ontology and metadata repository provides an ontological and lexical framework for describing ASDC data holdings, as well as support for deductive reasoning and querying capabilities. The aforementioned RDF format was chosen to represent this information. An intuitive web-based user interface was developed, allowing users to quickly sort through and analyze relevant portions of the ontology in order to locate the desired data.

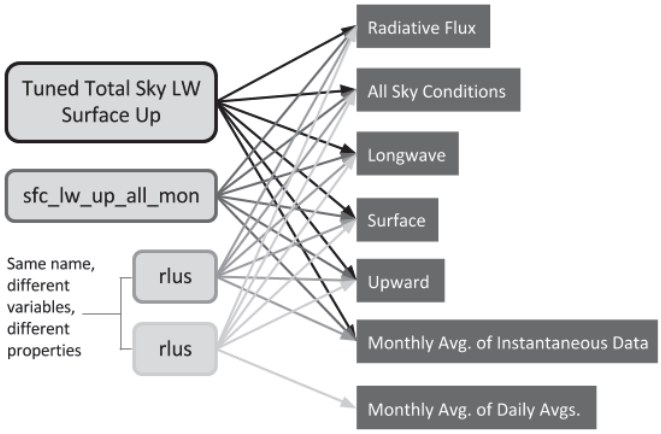


Fig. 1: Variables' semantic content. Variable names on the left; semantic content on the right. Note that differently named variables can have identical properties—i.e. "Tune Total Surface Sky LW Surface Up", "sfc_lw_up_all_mon", "rlus"—while identically named variables can have different properties—i.e. "rlus" and "rlus".

3.1 Concept Modeling

ODISEES was, in large part, developed around the idea that recognition is generally faster, easier, and simpler than recall. As such, the model itself is not overly concerned with terminology or nomenclature, and instead puts greater emphasis on the semantics and relationships of concepts. The ontology was structured with the intention of merging a controlled vocabulary with useful deductive reasoning systems that can support semi-intelligent applications. Controlled vocabularies can, however, be overly restrictive and cumbersome for some types of applications, such as text-based searches, insofar as they often require significant effort on the user's part to either memorize a lot of terms or spend a lot of time looking up the correct term. Instead of using this kind of terminology-centric approach, the ODISEES ontology focuses on modeling the domain concepts that give meaning to terms, and treats these terms as first-class objects that refer to the concepts in the domain.

For example, a data variable is described in the ontology as a set of relationships between it and other domain objects. A Radiative Flux variable will have a relationship to a certain

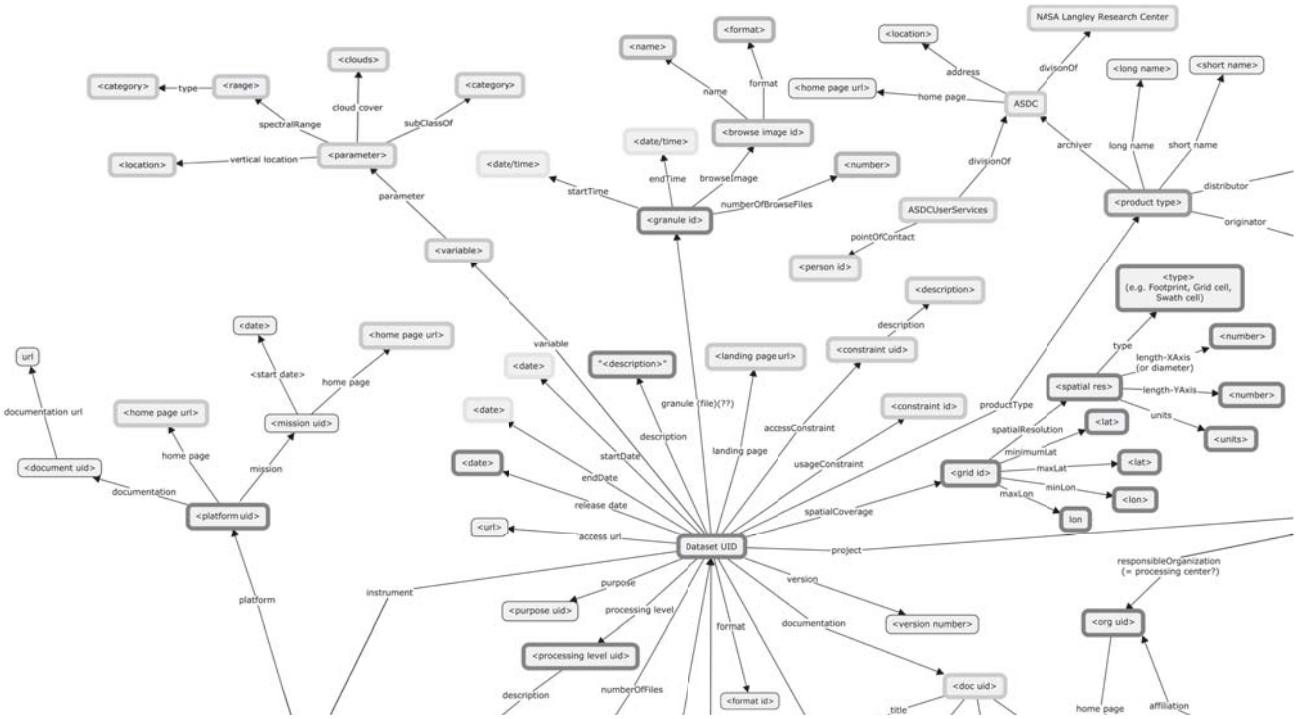


Fig. 2: Concept models for data set ("Dataset UID") and variable. The data set model is at the bottom-middle, and the variable model is at the top-left. Note the line connecting the two models, which indicates that the variable class is an attribute of the data set class. This figure shows only a portion of the overall concept map, which contains concept models for all classes within the ODISEES ontology—hence the other models partially shown alongside the data set and variable models.

wavelength or spectral range. The relationship is represented as a binary function that takes a particular atmospheric radiation variable, such as "CERES SW TOA Flux Upwards", and a particular spectral range, such as "shortwave", as arguments. Each variable's set of attributes is determined by analyzing the variable itself, the data sets in which it is included, metadata about the data set, and reviewing product documentation as well as consulting with domain experts. Taken together, the set of relationships describe the variable in a machine-readable format with sufficient detail to allow a scientist to assess, with reasonable precision, the essential characteristics of the observation or model output represented. This representation provides maximum flexibility and extensibility in describing Earth Science data and model outputs because, at any point, new domain objects can be introduced, new relations can be defined, and new relationships among variables and other domain objects can be asserted, without requiring any changes to the underlying data model. As a result, variables, or indeed any object in the model, can be identified and evaluated strictly in terms of their defining attributes and without regard for the naming convention that may have been followed in labeling it. Figure 1 provides an illustration of how this semantic content is attached to variable names.

Each attribute of a given object in the ODISEES model is itself an object and has a label or name. Variable objects are themselves attributes of the data set objects. Figure 2 shows the concept models for both the variable class and the data set class.

A primary strength of this kind of model is that even if the application requirements change, the model can be easily adapted to accommodate these changes. For the ODISEES search application, we want some of the attributes of a data set to be inherited by the variables within that data set. Similarly, if characteristics of the remote sensing instrument used to create the data set have implications for the data it produces, we want the data set or the variables within it to inherit whatever attributes are implied. For example, if a sensor is calibrated to measure radiation in a particular spectral range, the spectral range associated with a data variable produced by that instrument can be automatically inferred from the fact that it was produced by that instrument. The inferential capabilities of our logic-based model were used to materialize many of the assertions we wanted to drive the application.

3.2 Variable Discovery and Comparison

As mentioned in the beginning of Section 3, the core purpose of ODISEES is not just to represent or conceptualize

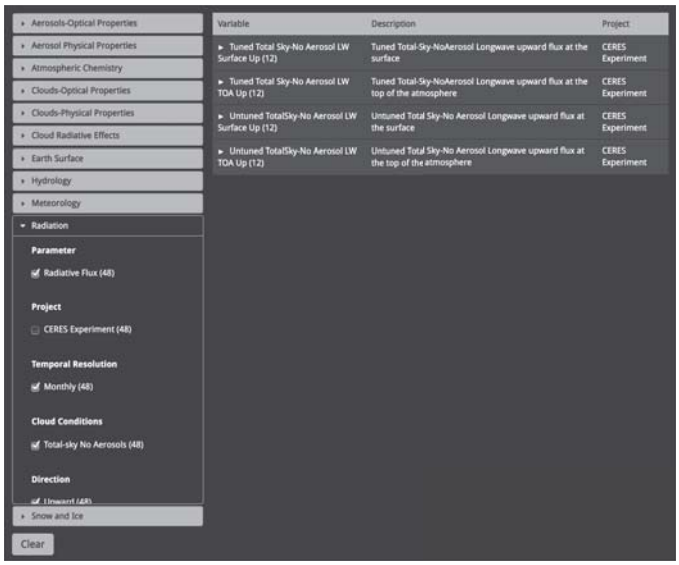


Fig. 3: Attribute filter selection with the ODISEES user interface.

Variable Name	Tuned Total Sky-No Aerosol LW Surface Up	sfc_lw_up_clr_mon
Cloud Conditions	Total-sky No Aerosols	Clear Sky
Data Set	CER_SYN1deg-M3Hour_Terra-Aqua-MODIS_Edition3A-Global	CERES EBAF Surface Edition 2.7
Data Source	Satellite Observation	Satellite Observation
Description	Tuned Total-Sky-NoAerosol Longwave upward flux at the surface	undefined
Dimensions	Statistic Type	time lon Latitude
Direction	Upward	Upward
format	HDF 4	NetCDF
grid type	undefined	Equal Angle Grid
Instrument	CERES FM3 CERES FM2 CERES FM4 CERES FM1	CERES FM3 CERES FM2 CERES FM4 CERES FM1
Method	Monthly Average of 3 Hour Intervals (8 per day)	undefined
Parameter	Radiative Flux	Radiative Flux
Project	CERES Experiment	CERES Experiment
Spatial Coverage	Global	Global
Spatial Resolution (Horizontal)	Earth's Surface (WGS-84 Earth Model)	1° lat x 1° lon
Spectral Range	Longwave (from 4 µm)	Longwave (from 4 µm)
Temporal Resolution	1 Month	1 Month
Unit of Measure	W/m^2	W/m^2
Vertical Location	Earth's Surface (Land and/or Water)	Earth's Surface (Land and/or Water)
Vertical Location Details	Earth's Surface (WGS-84 Earth Model)	Earth's Surface (WGS-84 Earth Model)
Wavelength Details	5-100 µm (LW)	5-100 µm (LW)

Fig. 4: Variable comparison with the ODISEES user interface.

information, but rather to make information more discoverable and thereby usable. More specifically, ODISEES is designed to leverage the concept modeling system described in Section 3.1 to enable users to more easily and effectively discover and compare variables from many distinct data sets. To this end, ODISEES uses semantic content attached to individual variables. This content is encoded as a set of RDF assertions that uniquely describes each variable. We say "uniquely" because, even though variables may have many attributes in common, the complete set of assertions that describes any given individual variable is unique to that variable.

The set of RDF assertions that describe the data variables are used as filters that let users specify criteria and thereby narrow their search to a set of results that satisfy all and only those criteria. Figure 3 shows a set of filters and the variables which satisfy the selected options as they appear in the ODISEES user interface.

Once the desired variables are selected, users can generate a comparison table, which displays each variable's respective semantic content items and highlights differences between them to aid and inform the comparison process. Figure 4 shows this comparison feature in the ODISES user interface.

3.3 Results

The ODISEES beta version is currently deployed and maintained at the ASDC, and is accessible to the public at odisees.larc.nasa.gov. We are actively developing improvements to the search capabilities and increasing the amount of searchable data in the ODISEES repository. There are currently 91 data sets represented in the ontology, and new ones are being added regularly. The ontology is still relatively small—400,931 RDF triples—but it's expected to grow significantly over the next two years. Initial user feedback has been generally positive, with many users attesting to the present and future usefulness of the tool. The test user group is composed of developers and researchers from NASA, NOAA, the EPA, multiple universities, and several other organizations.

4. Discussion

The ODISEES project discussed in this paper presents an adaptable, modular solution to some challenges posed by the extreme abundance of closely related data. Furthermore, it demonstrates the potential usefulness of ontology-based applications in solving issues posed by the growing overabundance of data.

Several features, including text search, simple data subsetting with the Open-source Project for a Network Data Access Protocol (OPeNDAP), and various improvements to the web interface, are planned for future releases. Additionally, development has begun on the Ontology-based Metadata Portal for Unified Semantics (OlyMPUS) [10], a metadata ingest system which leverages the same ontological structure as ODISEES. The ODISEES-OlyMPUS end-to-end system will support both data consumers and data providers, enabling the latter to register their data sets and provision them with the semantically rich metadata that drives ODISEES' data discovery and access service for data consumers.

5. Acknowledgments

The authors would like to thank the Atmospheric Data Center at NASA's Langley Research Center for their ongoing support.

References

[1] ASDC, "About the Atmospheric Science Data Center," 2015. [Online]. Available: https://eosweb.larc.nasa.gov/more-about-asdc.

[2] Cyc, "Cyc Homepage," 2015. [Online]. Available: http://www.cyc.com/.

[3] D. Lenat, S. Laningham, "Doug Lenat on Cyc, a Truly Semantic Web and Artificial Intelligence AI," 2008. [Online]. Available: http://www.ibm.com/developerworks/podcast/dwi/cm-int091608txt.html.

[4] J. Friedman, "The sole contender for ai" in *Harvard Science Review*, 2003.

[5] D. Lenat, G. Miller, and T. Yokoi, "Cyc, wondernet, and edr: critiques and responses" in *Communications of the ACM*, vol. 38, no. 11, 2003.

[6] J. Taubarer, "AI Founder Blasts Modern Research," 2003. [Online]. Available: http://archive.wired.com/science/discoveries/news/2003/05/58714?currentPage=all.

[7] D. Allemang, J. Hendler, *Semantic Web for the Working Ontologist*. Elsevier, 2011.

[8] J. Taubarer, "What is RDF?," 2006. [Online]. Available: http://www.xml.com/pub/a/2001/01/24/rdf.html.

[9] T. Gruber, "Ontology" in *Encyclopedia of Database Systems*, 2009.

[10] J. Gleason, E. Huffer, A. Ross, P. McInerey, P. Mehrotra, P. Rinsland, "Ontology-based metadata portal for unified semantics (olympus)", Proposal submitted in response to National Aeronautics and Space Administration (NASA) Research Announcement (NRA) Advanced Information Systems Technology Program, 2014.

Index Optimization with KNN considering Similarities among Features

Taeho Jo
Department of Computer and Information Engineering, Inha University, Incheon, South Korea

Abstract— *In this research, we propose that the K Nearest Neighbor should be for the index optimization, considering the feature similarities. In the reality, the dependencies and the relations among features always exist; texts which are given as features for encoding words into numerical vectors have similarities with others. In this research, we define the similarity measure which considers the feature values and the features, interpret the index optimization into the classification task where each word is classified into expansion, reservation, or removal, and use the measure for modifying the K Nearest Neighbor as the approach to the task. As the benefits from this research, we obtain the potential possibility of encoding words into their more compact representations and the improved discriminations among even sparse numerical vectors. Therefore, the goal of this research is to implement the index optimization systems with the benefits.*

Keywords: Word Categorization, Feature Similarity

1. Introduction

The index optimization refers to the process of adjusting a list of index terms by adding more similar words, reserving some words, and removing irrelevant words, in order to maximize the information retrieval performance. The scope of this research is restricted to the classification task where each word is classified into the three categories: 'expansion', 'reservation', and 'removal'. We prepare the sample words which are labeled with one of the three categories and define the factors which influence on the classification. By learning the sample words, we construct the classification capacity and classify novice words which are given afterward as the input into one of the three categories. In this research, we assume that a supervised learning algorithm is used as the approach to the index optimization which is set as a classification task.

Let us mention some challenges which we try to solve in this research. The dependency among features exist clearly, so the Bayesian networks were previously proposed as a machine learning based approach, but it requires the complicated analysis among features for using it[1]. The requirement of many features for keeping the robustness in encoding words into numerical vectors is caused by the assumption of feature independences. Because of very little coverage of each feature, we cannot avoid the sparse distribution where zero values are dominant in each numerical vector with more than 95%[3]. Therefore, this research is intended to solve the problems by considering the feature similarity as well as the feature value one.

Let us mention what we propose in this research as its idea. In this research, we consider the both similarity measures, feature similarity and feature value similarity for computing the similarity between numerical vectors. This research interprets the index optimization into a classification task where a machine learning algorithm is applicable. The KNN (K Nearest Neighbor) is modified into the version which accommodates the both similarity measures, and applied to the index optimization task which is mapped into a classification task. Therefore, the goal of this research is to improve the index optimization performance by solving the above problems.

Let us mention the benefits which we expect from this research. Computing the similarity between words using the feature similarity as well as the feature value similarity opens potentially the way of reducing the dimensionality of numerical vectors. The addition of one more similarity measure cuts down the information loss for computing the semantic similarity between words. By considering the both kinds of similarity measures, we expect from this research to improve the discriminations among numerical vectors which tend to be sparse.Therefore, the goal of this research is to implement the index optimization systems with their better performance by obtaining the benefits.

This article is organized into the four sections. In Section **??**, we survey the relevant previous works. In Section 3, we describe in detail what we propose in this research. In Section 4, we mention the remaining tasks for doing the further research.

2. Previous Works

Let us survey the previous cases of encoding texts into structured forms for using the machine learning algorithms to text mining tasks. The three main problems, huge dimensionality, sparse distribution, and poor transparency, have existed inherently in encoding them into numerical vectors. In previous works, various schemes of preprocessing texts have been proposed, in order to solve the problems. In this survey, we focus on the process of encoding texts into alternative structured forms to numerical vectors. In other words, this section is intended to explore previous works on solutions to the problems.

Let us mention the popularity of encoding texts into numerical vectors, and the proposal and the application of string kernels as the solution to the above problems. In 2002, Sebastiani presented the numerical vectors are the standard representations of texts in applying the machine learning

algorithms to the text classifications [4]. In 2002, Lodhi et al. proposed the string kernel as a kernel function of raw texts in using the SVM (Support Vector Machine) to the text classification [5]. In 2004, Lesile et al. used the version of SVM which proposed by Lodhi et al. to the protein classification [6]. In 2004, Kate and Mooney used also the SVM version for classifying sentences by their meanings [7].

It was proposed that texts are encoded into tables instead of numerical vectors, as the solutions to the above problems. In 2008, Jo and Cho proposed the table matching algorithm as the approach to text classification [8]. In 2008, Jo applied also his proposed approach to the text clustering, as well as the text categorization [12]. In 2011, Jo described as the technique of automatic text classification in his patent document [10]. In 2015, Jo improved the table matching algorithm into its more stable version [11].

Previously, it was proposed that texts should be encoded into string vectors as other structured forms. In 2008, Jo modified the k means algorithm into the version which processes string vectors as the approach to the text clustering[12]. In 2010, Jo modified the two supervised learning algorithms, the KNN and the SVM, into the version as the improved approaches to the text classification [13]. In 2010, Jo proposed the unsupervised neural networks, called Neural Text Self Organizer, which receives the string vector as its input data [14]. In 2010, Jo applied the supervised neural networks, called Neural Text Categorizer, which gets a string vector as its input, as the approach to the text classification [15].

The above previous works proposed the string kernel as the kernel function of raw texts in the SVM, and tables and string vectors as representations of texts, in order to solve the problems. Because the string kernel takes very much computation time for computing their values, it was used for processing short strings or sentences rather than texts. In the previous works on encoding texts into tables, only table matching algorithm was proposed; there is no attempt to modify the machine algorithms into their table based version. In the previous works on encoding texts into string vectors, only frequency was considered for defining features of string vectors. Texts which are used as features of numerical vectors which represent words have their semantic similarities among them, so the similarities will be used for processing sparse numerical vectors, in this research.

3. Proposed Approach

This section is concerned with modifying the AHC (Agglomerative Hierarchical Clustering) algorithm into the version which considers the similarities among features as well as feature values, and it consists of the three sections. In Section 3.1, we describe the process of encoding words into numerical vectors. In Section 3.2, we do formally the proposed scheme of computing the similarity between two numerical vectors. In Section **??**, we mention the proposed version of AHC algorithm which considers the similarity among features as the approach to word clustering. Therefore, this article is intended to describe in detail the modified version of KNN algorithm and its application to the word clustering.

3.1 Word Encoding

This subsection is concerned with the process of encoding words into numerical vectors. Previously, texts each of which is consists of paragraphs were encoded into numerical vectors whose attributes are words. In this research, we attempt to encode words into numerical vectors whose attributes are text identifiers which include them. Encoding of words and texts into numerical vectors looks reverse to each other. In this Section, we describe in detail the process of mapping words into numerical vectors, instead of texts.

In the first step of word encoding, a word-document matrix is constructed automatically from a text collection called corpus. In the corpus, each text is indexed into a list of words. For each word, we compute and assign its weight which is called TF-IDF (Term Frequency-Inverse Document Frequency) weight [2], by equation (1),

$$w_i = TF_i(\log_2 N - \log_2 DF_i + 1) \tag{1}$$

where TF_i is the total frequency in the given text, DF_i is the total number of documents including the word, and N is the total number of documents in the corpus. The word-document matrix consists of TF-IDF weights as relations between a word and a document computed by equation (1). Note that the matrix is a very huge one which consists at least of several thousands of words and documents.

Let us consider the criterion of selecting text identifiers as features, given labeled sampled words and a text collection. We may set a portion of each text in the given sample words as a criteria for selecting features. We may use the total frequency of the sample words in each text as a selection criterion. However, in this research, we decided the total TF-IDF (Term Frequency and Inverse Document Frequency) which is computed by equation (1) as the criterion. We may combine more than two criteria with each other for selecting features.

Once some texts are selected as attributes, we need to consider the schemes of defining a value to each attribute. To each attribute, we may assign a binary value indicating whether the word present in the text which is given as the attribute, or not. We may use the relative frequency of the word in each text which is an attribute as a feature value. The weight of word to each attribute which is computed by equation (1) may be used as a feature value. Therefore, the attributes values of a numerical vector which represent a word are relationships between the word and the texts which are selected as features.

The feature selection and the feature value assignment for encoding words into numerical vectors depend strongly on the given corpus. When changing the corpus, different texts are selected by different values of the selection criterion as features. Even if same features are selected, different feature values are assigned. Only addition or deletion of texts

in the given corpus may influence on the feature selection and the assignment of feature values. In order to avoid the dependency, we may consider the word net or the dictionary as alternatives to the corpus.

3.2 Feature Similarity

This subsection is concerned with the scheme of computing the similarity between numerical vectors as illustrated in Figure 1. In this research, we call the traditional similarity measures such as cosine similarity and Euclidean distance feature value similarities where consider only feature values for computing it. In this research, we consider the feature similarity as well as the feature value similarity for computing it as the similarity measure which is specialized for text mining tasks. The numerical vectors which represent texts or words tend to be strongly sparse; only feature value similarity becomes easily fragile to the tendency. Therefore, in this subsection, as the solution to the problem, we describe the proposed scheme of computing the similarity between numerical vectors.

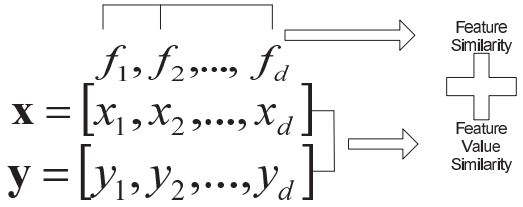


Fig. 1
THE COMBINATION OF FEATURE AND FEATURE VALUE SIMILARITY

Text identifiers are given as features for encoding words into numerical vectors. Texts are dependent on others rather than independent ones which are assumed in the traditional classifiers, especially in Naive Bayes [1]. Previously, various schemes of computing the semantic similarity between texts were developed [2]. We need to assign nonzero similarity between two numerical vectors where non-zero elements are given to different features with their high similarity. It is expected to improve the discriminations among sparse vectors by considering the similarity among features.

We may build the similarity matrix among features automatically from a corpus. From the corpus, we extract easily a list of text identifiers. We compute the similarity between two texts by equation (2),

$$s_{ij} = sim(d_i, d_j) = \frac{2 \times tf(d_i, d_j)}{tf(d_i) + tf(d_j)} \tag{2}$$

where $tf(d_i, d_j)$ is the number of words which are shared by both texts, d_i and d_j, and $tf(d_i)$ is the number of words which are included in the text, d_i. We build the similarity matrix which is consists of similarities between text identifiers given as features as follows:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1d} \\ s_{21} & s_{22} & \cdots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \cdots & s_{dd} \end{pmatrix}.$$

The rows and columns in the above matrix,S, correspond to the d text identifiers which are selected as the features.

The texts, $d_1, d_2, ..., d_d$ are given as the features, and the two words, t_1 and t_2 are encoded into the two numerical vectors as follows:

$$t_1 = [w_{11}, w_{12}, ..., w_{1d}]$$
$$t_2 = [w_{21}, w_{22}, ..., w_{2d}].$$

The features,$d_1, d_2, ..., d_d$ are defined through the process which was described in Section 3.1. We construct the d by d matrix as the similarity matrix of features by the process mentioned above. The similarity between the two vectors are computed with the assumption of availability of the feature similarities, by equation (3),

$$sim(t_1, t_2) = \frac{\sum_{i=1}^{d} \sum_{j=1}^{d} s_{ij} w_{1i} w_{2j}}{d \cdot \|t_1\| \cdot \|t_2\|} \tag{3}$$

where $\|t_1\| = \sqrt{\sum_{i=1}^{d} w_{1i}^2}$ and $\|t_2\| = \sqrt{\sum_{i=1}^{d} w_{2i}^2}$. We get the value of s_{ij} by equation (2).

The proposed scheme of computing the similarity by equation (3) has the higher complexity as payment for obtaining the more discrimination among sparse vectors. Let us assume that two d dimensional numerical vectors are given as the input for computing the similarity between them. It takes only linear complexity, $O(d)$, to compute the cosine similarity as the traditional one. However, in the proposed scheme takes the quadratic complexity, $O(d^2)$. We may reduce the complexity by computing similarities of some pairs of features, instead of all.

3.3 Proposed Version of KNN

This section is concerned with the version of K Nearest Neighbor which considers both the feature similarity and the feature value one. The sample words are encoded into numerical vectors whose features are texts by the scheme which was described in section 3.1. The novice word is given as the classification target, and it is also encoded into a numerical vector. Its similarities with the sample words are computed by equation (3) for selecting nearest neighbors, in the proposed version. Therefore, in order to provide the detail algorithm, we describe the proposed KNN version, together with the traditional one.

The traditional KNN version is illustrated in Figure 2. The sample words which are labeled with the positive class or the negative class are encoded into numerical vectors. The similarities of the numerical vector which represents a novice word with those representing sample words are computed using the Euclidean distance or the cosine similarity. The

k most similar sample words are selected as the k nearest neighbors and the label of the novice entity is decided by voting their labels. However, note that the traditional KNN version is very fragile in computing the similarity between very sparse numerical vectors.

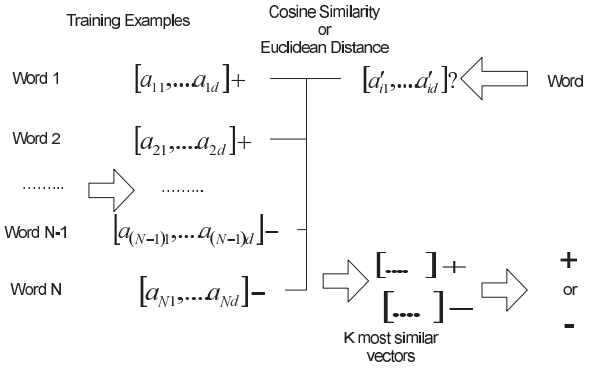


Fig. 2
THE TRADITIONAL VERSION OF KNN

The proposed KNN version is illustrated in Figure 3. Like the traditional version, a word is given as an input and it is encoded into a numerical vector. The similarities of the novice word with the sample ones are computed by equation (3) which was presented in section 3.2. Like the traditional version, k most similar samples are selected as the nearest neighbors, and the label of the novice is decided by voting their labels. The scheme of computing the similarity between numerical vectors is the essential difference between the two versions.

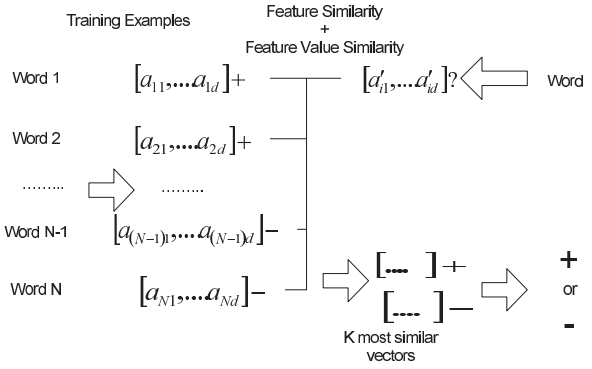


Fig. 3
THE PROPOSED VERSION OF KNN

We may derive some variants from the proposed KNN version. We may assign different weights to selected neighbors instead of identical ones: the highest weights to the first nearest neighbor and the lowest weight to the last one. Instead of a fixed number of nearest neighbors, we select any number of training examples within a hyper-sphere whose center is the given novice example as neighbors. The categorical scores are computed proportionally to similarities with training examples, instead of selecting nearest neighbors. We may also consider the variants where more than two variants are combined with each other.

Let us compare the both KNN versions with each other. In computing the similarity between two numerical vectors, the traditional version uses the Euclidean distance or cosine similarity mainly, whereas the proposed one uses the equation (3). Both versions are common in selecting k nearest neighbors and classifying a novice item by voting the labels of them. However, the proposed version is more tolerant to sparse numerical vectors in computing the similarities among them than the traditional version.

3.4 The Application to Index Optimization

This section is concerned with the scheme of applying the proposed KNN version which was described in Section 3.3 to the index optimization task. Before doing so, we need to transform the task into one where machine learning algorithms are applicable as the flexible and adaptive models. We prepare the words which are labeled with ‘expansion’, ‘inclusion’ or ‘removal’ as the sample data. The words are encoded into numerical vectors by the scheme which was described in Section **??**. Therefore, in this section, we describe the process of extracting words which belong to the two categories, ‘expansion’ and ‘inclusion’, from texts automatically using the proposed KNN with the view of the index optimization into a classification task.

In this research, the index optimization is viewed into a classification task, as shown in Figure 4. A text is given as the input, and a list of words is extracted by indexing the text. Each word is classified by the classifier into one of the three categories: ‘expansion’, ‘inclusion’, or, ‘removal’. In the task, the text is mapped into words which are classified with ’expansion’ or ’inclusion’. The similar words to one labeled with ’expansion’ will be added from external sources.

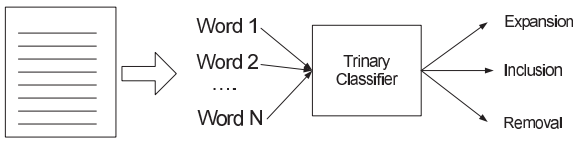


Fig. 4
MAPPING OF INDEX OPTIMIZATION INTO CLASSIFICATION TASK

We need to prepare sample words which are labeled with one of the three categories, before classifying a novice one or ones. A text collection is segmented into sub-collections of content based similar words which are called domains, manually or automatically. We prepare sample words which are labeled manually, domain by domain. To each domain, we assign and train a classifier with the words in the corresponding sub-collection. When a text is given as the input, the classifier which corresponds to the most similar

Let us consider the process where an article is given as the input and a list of essential words is extracted as the output.

We nominate the classifier which corresponds to the subgroup which is closest to the given article with respect to its content. A list of words is extracted by indexing the article, and each word is encoded. The words are classified by the nominated classifier into one of the three categories, and we select ones which are labeled with 'expansion' or 'reservation' as the optimized index. The addition of external words which are semantically similar as ones labeled with 'expansion' is set as the subsequent task.

Even if the index optimization is viewed into an instance of word categorization, it needs to be distinguished from the topic based word categorization. The word categorization is given as a single multiple classification or multiple binary classifications, whereas the index optimization is done as a single triary classification or three binary classification tasks. In the word categorization, each word is classified semantically into one or some of the predefined topics, whereas in the index optimization, it is classified one of the three actions. In the word categorization, each word is classified by its meaning, whereas in the index optimization, it is classified by its importance to the given text. In the word categorization, when the given task is decomposed into binary classification tasks, a classifier is assigned to each topic, whereas, in the index optimization, a classifier is done to each domain.

4. Conclusion

Let us mention the remaining tasks for doing the further research. We need to validate the proposed approach in specific domains such as medicine, engineering, and economics, as well as in generic domains such as ones of news articles. We may consider the computation of similarities among some main features rather than among all features for reducing the computation time. We try to modify other machine learning algorithms such as Naive Bayes, Perceptrons, and SVM (Support Vector Machine) based on both kinds of similarities. By adopting the proposed approach, we may implement the word clustering system as a real program.

References

[1] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, Addison-Wesley, 2011.
[3] T. Jo, "The Implementation of Dynamic Document Organization using Text Categorization and Text Clustering" PhD Dissertation, University of Ottawa, Ottawa, Canada, 2006.
[4] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Survey*, Vol. 34, pp. 1-47, 2002.
[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification with String Kernels", *Journal of Machine Learning Research*, Vol. 2, pp. 419-444, 2002.
[6] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch String Kernels for Discriminative Protein Classification", *Bioinformatics*, Vol. 20, pp. 467-476, 2004.
[7] R. J. Kate and R. J. Mooney, "Using String Kernels for Learning Semantic Parsers", in *Proc. ICCL '06*, 2006, pp. 913-920.
[8] T. Jo and D. Cho, "Index based Approach for Text Categorization", *International Journal of Mathematics and Computers in Simulation*, Vol. 2, 2008, pp. 127-132.
[9] T. Jo, "Single Pass Algorithm for Text Clustering by Encoding Documents into Tables", *Journal of Korea Multimedia Society*, Vol. 11, 2008, pp. 1749-1757.
[10] T. Jo, "Device and Method for Categorizing Electronic Document Automatically", South Korean Patent 10-1071495, 2011.
[11] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", *Soft Computing*, Vol. 19, 2015, pp. 849-849.
[12] T. Jo, "Inverted Index based Modified Version of K-Means Algorithm for Text Clustering", *Journal of Information Processing Systems*, Vol. 4, 2008, pp. 67-76.
[13] T. Jo, "Representationof Texts into String Vectors for Text Categorization", *Journal of Computing Science and Engineering*, Vol. 4, 2010, pp. 110-127.
[14] T. Jo, "NTSO (Neural Text Self Organizer): A New Neural Network for Text Clustering", *Journal of Network Technology*, Vol. 1, 2010, pp. 31-43.
[15] T. Jo, "NTC (Neural Text Categorizer): Neural Network for Text Categorization", *International Journal of Information Studies*, Vol 2, 2010, pp. 83-96.

Classification based Filtering for Personalized Information Retrieval

Sachintha Pitigala[1], Cen Li[2]
[1] Center for Computational Sciences, MTSU, Murfreesboro, TN, USA
[2] Department of Computer Science, MTSU, Murfreesboro, TN, USA
*spp2k@mtmail.mtsu.edu, Cen.Li@mtsu.edu

***Abstract*— PubMed keyword based search often results in many citations not directly relevant to the user information need. Personalized Information Retrieval (PIR) systems aim to improve the quality of the retrieval results by letting users supply more information than keywords. There are two main problems relate to current PIR systems developed for PubMed: (1) requiring the user to supply a large number of citations directly relevant to search topic, and (2) produces too many search results with high false positive. This paper describes a Classification based multi-stage Filtering (ClaF) approach to address these problems. A small set of citations relevant to the information need is needed from the user. The system automatically finds similar citations to the inputs and builds a larger training set. This training set is used to train multiple text classifiers, each with a different classification scheme. The trained text classifiers are used in a Multi-stage filtering process to find the relevant citations to the user information need. Results show the proposed ClaF system is feasible and produces good retrieval results.**

Keywords: Information Retrieval, Personalized Information Retrieval, Text Classification, PMRA, PubMed.

I. INTRODUCTION

SCIENTIFIC literature databases had an exponential growth over the past decade. Google Scholar [1], PubMed [2], The SAO/NASA Astrophysics Data System [3] and CiteSeerX [4] are some of the popular citation databases on the internet. These online databases open a new way of accessing and searching for the information for the scientific community.

PubMed is the largest literature database in the biomedicine field. PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medcine (NLM) [5]. It contains over 24 million biomedicine and related citations covering over 5000 journals ([2], [5]). Given a keyword based user query, PubMed typically returns a large number of citations relevant to the search query. For example, over one-third of PubMed queries returned 100 or more citations [6]. Sifting through these citations to locate the ones that represent the most relevant articles for one's query can be a time consuming process. It is desirable to have search tools that are capable of capturing each user's unique research interest and returning a smaller set of citations of the truly relevant articles from a large literature databases such as PubMed. These types of search tools are referred as Personalized Information Retrieval (PIR) Systems.

For the traditional Information Retrieval (IR) systems, user information needs are provided as user queries consisting of keyword terms. For PIR systems, the unique interests of a user's information need are better captured with the use of additional information provided by the user. Currently, PIR systems can be divided into two main categories based on the way it gathers the user interest. The first category of the PIR systems gathers the user information and interest explicitly from the user ([7], [8], [9], [10], [11]). The second category of PIR the systems gathers its user personalized information implicitly, for example in terms of the click-through links in the search history ([12], [13], [14]). This research focuses on developing a PIR system for the PubMed based on user explicit information.

Many explicit PIR systems allow users to provide additional information about their query through an advanced search option where the user may explicitly enter the area of interest, publication period, journals or authors of interest, along with the query terms ([7], [8]). These additional information further filters the search output, thus reduces the size of the search output.

Yet, it is possible to get more explicit information about the user's query intent in order to deliver more personalized results. For example, explicit PIR systems allow the users to enter a text paragraph to explain his/her information need, or input paragraphs or the abstract of a research article. eTBLAST [9] is an explicit PIR tool for PubMed built based on free text inputs. The inputs provide more information to the search tool than the keywords. It produces better results compared to the traditional keyword based method.

MedlineRanker [10] and MScanner [11] are explicit PIR systems for PubMed that take as input a set of citations that are deemed relevant to a user's information need. The systems derive the information needs from this set and searches for the relevant citations best matching the input. The focus of the systems is on ranking the search results based on the input citations. They do not directly focus on reducing the search output size from PubMed. Both systems require a user to provide at least 100 citations highly relevant to the user interest in order to get reasonable search results. This requirement is unrealistic in many situations.

*Corrosponding Author: Sachintha Pitigala. Email: spp2k@mtmail.mtsu.edu

The goal of this research is to build a PIR System for PubMed that is capable of delivering highly relevant search results, reducing the search output size by limiting irrelevant citations in the search output, and only requires the users to input a small set of citations of the relevant articles. In this study, the proposed PIR system is referred as Classification based Filtering (ClaF) system.

To evaluate the performance of the ClaF system, TREC 2005 dataset [15] is used. This dataset consists of 50 information needs from real biomedicine researchers. Each information need in TREC contains a document pool and each document in the pool is labeled as Definitely Relevant (DR), Possibly Relevant (PR) or Not Relevant (NR) [15].

The ClaF system takes a set of PubMed citations as user input. The input citations represent the user research interest or information need. We call this citation set the user *seed* documents. *Seed* documents typically consist of 5 to 20 citations carefully chosen by the user. It has been reported that learning text classifiers based on a small training data is difficult ([16], [17]). To better illustrate this difficulty, result from a simple experiment is discussed here. To form the training data, first, five documents were randomly selected from the combined DR and PR set of an information need. Then five more documents were randomly selected from the NR set of the same information need. Naive Bayes text classifier is trained using the 10 documents. Table 1 shows the average classification results (averaged over 10 random runs) of the classifiers trained for five different information needs. It is clear that the classification accuracies of the text classifiers are extremely low and it is making many false positive classifications.

In order for the PIR system to be effective in retrieving relevant citations based only on a small number of citations from the user, the system should be able to:

1. Increase the size of the training data based solely on the user *seed* citations while maintaining the quality of the training data;
2. Reduce the false positive classifications
3. Rank the final search output efficiently and effectively.

To achieve the first goal, a method based on PubMed Related Articles (PMRA) [18] and Cosine Similarity [19] is developed. A Text Classification based multi-stage filtering model is used to reduce the number of false positive classifications. Finally, cosine similarity measure and the *seed* citations are used to rank the final predicted relevant documents. Experimental results from 10 different information-needs from the TREC 2005 dataset show that the system produces reliable search results for the given information need.

The rest of the paper is organized as the following: Section 2 discusses the PMRA and the cosine similarity measure and text classification methods used in this study. Section 3 presents the proposed ClaF system, TREC 2005 genomic dataset and preprocessing steps. Section 4 describes the experiment procedure and experimental results of the ClaF system. Section 5 discusses the conclusions about the study and presents the future research directions.

Table 1: Accuracy of Naive Bayes classifiers for five information needs from the TREC 2005 dataset. In this experiment, the training set contains 5 relevant and 5 non-relevant citations from the information need.

Topic ID (Information Need)	Precision	#Articles classified as positive	# Articles correctly classified as positive	# Actual positive articles in dataset
117	0.06617	12360	685	704
146	0.02891	16013	420	432
120	0.02811	13139	318	340
114	0.02905	13903	354	374
126	0.02487	13856	284	302

II. Background

To increase the size of the training data based on *seed* documents supplied by a user, a similarity-based approach is developed to find citations from the entire database that are similar to the *seed* citations. Given the size of the PubMed database, to perform a real time similarity computation between each of the *seed* citation and every citation in the database is generally not practical. Therefore, this study uses the PMRA feature [18] to build a small *Target Set*, based on which a larger training data is formed.

A. PubMed Related Articles (PMRA) feature

The PMRA feature computes the similarity between pair-wise citations in the database. The relevancy between two citations is calculated using the words they have in common, with citation length adjustment. Words from title, abstract and Medical Subject Headings (MeSH) terms are used to represent a citation in this algorithm. PubMed related citations are calculated using the entire PubMed database for given citation. This process takes several days to complete [20]. Therefore, PubMed related citation list for any given citation is pre-calculated and sorted in the PubMed. The most relevant citations for any given citation, called the PMRA list, are stored in PubMed database. The PMRA list is a useful feature in PubMed. A PubMed log analysis showed that a fifth of PubMed searches invoke the PubMed related articles (citations), suggested by the PMRA list, at-least once [21].

In our system, the PMRA lists of the user *seeds* are combined to form a *Target Set*. This *Target Set* is then used to find more positive training example for text classification. Cosine Similarity measure is chosen as the similarity measure to find documents similar to the ones in the *seed* set.

B. Cosine Similarity

Cosine similarity is heavily used in the information retrieval and text mining community. A previous study showed that cosine similarity and the overlap model out-performed many other similarity measures in the TREC dataset [22]. Cosine Similarity provides a simple and effective method to compute the similarity between articles by measuring the angle between the two vectors representing the two articles.

In this study, cosine similarity is used to compute the similarity between each citation in the *Target set* and the ones in the *seed* set. The candidate citations with the highest similarity values are added to the positive training example set.

Once the training data set is formed, ClaF system learns the text classifiers based on the training data. Text classification automatically assigns documents into one or more categories based on its content. Popular text classification approaches include the Naive Bayes (NB) classification [23], the Support Vector Machines (SVM) [23], the Rocchio method [24], the regression based models [25], the k-Nearest Neighbor (kNN) method [24], and the Neural Networks [25]. NB, kNN and SVM text classification approaches have been used in the ClaF system.

The following sections briefly explain the theory behind the kNN text classification, the Naive Bayes classification, and Support Vector Machine (SVM) approach.

C. k-Nearest Neighbor (kNN) Text Classification

k-Nearest Neighbor (kNN) algorithm is also known as instance-based learning or lazy learning. The kNN algorithm does not have an explicit training step. During classification, it examines the class labels of k nearest neighbors that are the most similar to the test object, and classifies the test object with the majority label from its k neighbors. A similarity measure is used to find the k nearest neighbors from the training set. Cosine Similarity is used here to find the nearest neighbors. In this study, kNN training set consists of equal number of positive (relevant) and negative (non-relevant) training examples. We need to predefine a value for k in the kNN text classifier. In order to break ties in majority vote, an odd integer for k such as 1,3,5,7... is often used. The best value of k depends on the dataset.

D. Naive Bayes Text Classification

The Naive Bayes is a fast and robust text classification method. It is based on the *posterior* probability model derived using the Bayes theorem [23]. Given a document d, its probability of belonging to a class c is $P(c|d)$. The goal of the Naive Bayes classification is to find the optimal class for a given document, i.e., the class that gives the maximum posterior probability, $\hat{P}(c|d)$ [23]. This is expressed as:

$$C_{map} = \arg\max_{c \in C} \hat{P}(c|d) \qquad (1)$$

where, C_{map} is the class with the maximum posterior probability, $c \in \{c_1, c_2, c_3, \ldots . c_n\} = C$ is the set of class labels and d is the given document. Then, applying the Bayes theorem and the Naive Bayes conditional independence assumption Equation 1 can be re-written as:

$$C_{map} = \arg\max_{c \in C} \prod_{i=1}^{nd} \hat{P}(t_i|c)\ \hat{P}(c) \qquad (2)$$

where $\{t_1, t_2, t_3, \ldots . t_{nd}\}$ is the set of terms in document d, and nd is the total number of terms in the document. During the training stage, the probabilities, $\hat{P}(c)$ and $\hat{P}(t_i|c)$, are estimated from the training data. $\hat{P}(c)$ is the prior probability of class c.

At the classification stage, given terms $\{t_1, t_2, t_3, \ldots . t_{nd}\}$ for a document d, Equation 2 is used to compute the posterior probability of the document for each possible class, $c \in C$. The class assigned to the document is the one having the highest posterior probability.

E. Support Vector Machines (SVM)

SVM is a popular and powerful algorithm for text classification and many other pattern recognition problems. It is originally a non-probabilistic binary classifier invented by Vapnik and his colleagues [26]. SVM method gives a formal explanation to find the optimal hyper-plane to separate data. Moreover, it finds the optimal hyperplane that maximizes the margin between two data regions. The data points on the marginal hyperplanes are called *support vectors*. If the initial data is not linearly separable in the feature space, SVM uses kernel functions to transform data into a higher dimensional feature space where a hyperplane exists to do the separation. In this study, the LIBSVM software [27] with WEKA [28] is used to perform SVM classification.

III. METHODOLOGY

The overall architecture of the proposed text Classification based multi-stage Filtering (ClaF) system is presented in Figure 1. ClaF requires a user to input a small set (e.g., 5 to 10) of citations. These citations represent the user information need and are referred as *seeds*. From the *seeds*, ClaF extracts the useful information for information retrieval. The following section presents the steps ClaF uses to extract the information from the *seeds*.

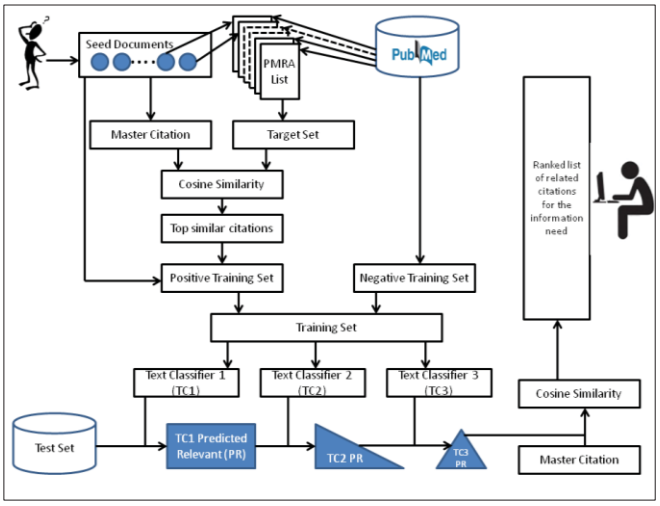


Figure 1: Overview of the ClaF system.

A. Data Preprocessing of user seeds

First, citation title, abstract and the Medical Subject Headings (MeSH) terms are extracted from each user *seed*. Information such as details about the author, affiliation data and journal information are ignored in this study. Then, the title, abstract text and the MeSH terms are tokenized into list of terms. From the document term list, stopwords [29] and

words containing only digits are removed. Next, stemming is applied to obtain a normalized term list for the document. Finally, the normalized terms from the title and MeSH terms with subheading qualifier are added again to the normalized term list to give more weight to those terms. Normalized term list from each user *seed* are used to build a *Master Citation* in the ClaF system. This *Master Citation* is used to represent the user information need.

B. Building the Master Citation

The set of user *seeds* collectively represents the user information need. Each single *seed* represents a segment of the user information need. Therefore, it is necessary to combine the *seeds* into a single citation to find the similar citations from the *Target Set*. This single citation is referred as the *Master Citation* in the system.

To form the *Master Citation*, first, a unique word list is created along with document frequencies and term frequencies using the normalized term list from each *seed*. Then, all the terms appearing in two or more *seeds* are added to the *Master Citation* term list. *Master Citation* is a unique representation of user information need. Next, *Master Citation* is used to build a larger training set for text classification in the ClaF system.

C. Expand the Training Set.

The experimental results given in the introduction section have shown the need to have more training examples in order to learn a more accurate text classifier. However, requesting a larger *seed* set from the user is not practical. Therefore, an automated procedure is needed to expand the training set based on the small set of user *seeds*.

To expand the training set, ClaF searches for documents that are the most similar to the *Master Citation* using the Cosine Similarity. To speed up this process, a *Target Set* of citations, e.g., a subset of the PubMed database, is formed from which potential documents are searched.

The PMRA lists [18] are used to build the *Target Set*. For each given citation in PubMed, its PMRA list is pre-calculated. To build the *Target Set*, first, the PMRA list for each user *seed* is retrieved. Then, *seed* PMRA lists are combined into a single citation list. This unique citation list is called as the *Target set*. Next, ClaF finds the documents similar to the *Master Citation* from this *Target set* using the Cosine Similarity. Finally, citations having the highest similarity to *Master Citations* are added to the training set. Together, these newly added citations and the user *seeds* form the positive (relevant) training examples. A similar size document set is randomly selected from the entire PubMed database and labeled as negative (irrelevant) training examples.

Next, a text classifier based multi-stage filtering takes place to gradually refining/reducing the set of citations classified/predicted as relevant.

D. Multi-stage Filtering using Text Classification

At the beginning of the filtering process, 3 classifiers are learned from the expanded training data set. In this study, Naive Bayes (NB), Support Vector Machines (SVM) and k-Nearest Neighbor (kNN) text classifiers are used as the three base text classifiers. The three learned classifiers are applied in 3 stages in refining and filtering of the retrieval results.

Stage 1 text classifier (TC1) is first used to classify the test set into two categories: relevant (positive) and irrelevant (negative). Test set can be the entire PubMed database or a subset of PubMed database selected by the user. For example, if the user is interested in only retrieving the relevant citations published in the last five years. Then, the test set includes only the citations published in those five years. The set of citations predicted as positive by TC1 is often quite large, including many false positives.

To remove the false positives from the retrieval results, citations classified as positive by TC1 undergoes two more classifications using stage 2 Text Classifier (TC2) and stage 3 Text Classifier (TC3) in a pipeline fashion. Only the citations classified as positive from the previous stage are fed into the next classification stage for further refinement.

ClaF uses three-stage text classifier based filtering to refine the set of retrieved citations. A different choice of the base classification scheme at each of the 3 stages can lead to a slightly different final retrieval results. We take a conservative approach in choosing the classification schemes: apply classification schemes having high recalls in the early stages of the filtering pipeline and apply classifiers that are most susceptible in incorrectly remove true positives in later stages in the filtering pipeline, i.e., to preserve the true positives in the retrieval results as much as possible.

This approach is different from the standard voting schemes used for classification, where the accuracy of the voting schemes doesn't depend on the order of the classifiers used. This approach is also different from the active learning methods. While most active learning methods focus on improving the classification accuracy by incrementally improving the training data, in ClaF, the training data is improved just once through expansion. All the text classifiers are trained using the same expanded training data. After that, ClaF focuses on reducing the false-positives in the search output rather than improving the accuracy of the text classifier.

The three-stage filtering method may be generalized into filtering pipeline with more or less stages, i.e., multi-stage filtering. For example, one may use two-stage or four-stage filtering with two or four classifiers respectively. Classification schemes other than Naive Bayes, kNN, and SVM may be used in each stage of the process. The conservative approach should be used to order the classifiers in the filtering stages.

E. Ranking the Final Output

The classification results from TC3 represent a much-improved set of highly relevant citations to the user information need. However, it may still contain some of the false-positives. As the final step, ClaF ranks the resulting set of citations based on the Cosine Similarity of each against the *Master Citation*. The top ranked citations are presented as the final retrieval results.

IV. Results and Discussion

The ClaF methodology is tested and validated using the TREC 2005 ad hoc retrieval task dataset [15]. It contains 50 information needs (topics) from the real biologists. The entire document collection for the 50 topics contains 34,633 unique PubMed citations. Each information need (topic) has a corresponding set of labeled citations ranges between 290 and 1356 [15]. Expert biologists have labeled each citation as to whether or not it is relevant to the information need. The labels can be one of the following three: Definitely Relevant (DR), Possibly Relevant (PR) and Non Relevant (NR) for the given topic. The 10 topics having the highest number of relevant documents (definitely relevant and possibly relevant) are used in this study. Those topic numbers are 117, 146, 120, 114, 126, 109, 142, 111, 107 and 108. Next, the experimental procedure of this study is described.

A. Experiment Procedure

For each chosen information need (topic), ClaF uses the following steps to form the user information need and to retrieve the relevant citations:

- n ($n = \{5,10,15,20,25\}$) citations are randomly selected from the Definitely Relevant (DR) and Possibly Relevant (PR) set of the topic. Those n citations are labeled as the user *seeds* for the current topic;
- The PMRA lists for the *seeds* are retrieved and combined to form the *Target Set*;
- The *seeds* are pre-processed into terms and used to form the *Master Citation*;
- N ($N=50$) citations that have the highest cosine similarity to the *Master Citation* are computed from the *Target Set*; This set and the original n *seeds* form the positive training examples.
- Randomly select $n+N$ citations from the TREC 2005 genomic track dataset to form the negative examples;
- Train the three Text Classifiers using the expanded training data;
- Classify the TREC 2005 Genomic dataset using TC1.
- TC1 classifies a subset of citations as "relevant".
- Apply TC2 to refine and reduce the set of the "relevant" citations;
- Apply TC3 to further refine and reduce the set of the "relevant" citations from TC2;
- Compute and rank the Cosine similarities between the "relevant" citations from TC3 and the *Master Citation*.

Each experiment is repeated 10 times by randomly selecting *seeds* from the given information need. *Seed* set size (n) range from 5 to 25 with increment of 5. Results of the 10 information needs are presented in the next section.

B. Experimental Results

The experiments are designed to test the effectiveness of the ClaF system in terms of the effectiveness of each of its three main steps (1) expanding training set size by building *Target Set* and forming *Master Citation*, (2) multi-stage filter, and (3) ranking of the final output.

B.1 Improvements from Expanding the Training Data Set

If the size of the initial user *seeds* is 5 ($n = 5$), then the initial training data size is 10 with the negative training examples. After expanding the training set, a larger training set size of size 110 is obtained. This larger training set is used to train the three base classifiers. For kNN, three nearest neighbors are used to classify the new instances. Linear SVM method from the LibSVM [27] in WEKA [28] is used. The classification accuracies obtained using the expanded training data are compared against those of the original training data (*seed* only training data). Table 2 shows the average improvement of classification accuracies for the 10 information topics. Equation 3 calculates the improvement of accuracy for a topic. The average improvement over five different training sets is reported.

$$AI = \frac{(AETS - ASTS)}{ASTS} * 100\ \% \quad (3)$$

where, *AI = Accuracy Improvement*, *ASTS= Accuracy with the Seed only Training Set (Initial Training Set)* and *AETS= Accuracy with the Expanded Training Set*.

From Table 2, it is clear expanding the training set using the *Target Set* and *Master Citation* lead to a big improvement of the classification accuracies of base text classifier across all 10 information needs. The PMRA feature helps to build a high quality small *Target Set*, and Cosine Similarity is effective in computing citations having the highest similarity to the *Master Citation* from the *Target Set*.

Table 2: Improvement of Classification Accuracy for the three base classifiers using expanded training data

Topic ID	Average Improvement (%)		
	NB	3-NN	SVM
117	+ 61.20	+ 184.81	+ 60.92
146	+ 82.26	+ 155.81	+ 14.86
120	+ 150.31	+ 334.95	+ 61.13
114	+ 73.82	+ 158.35	+ 68.80
126	+ 150.31	+ 110.65	+ 11.75
109	+ 67.20	+ 82.81	+ 9.74
142	+ 182.62	+ 190.39	+ 150.09
111	+ 89.73	+ 225.86	+ 139.93
107	+ 66.96	+ 104.96	+ 24.58
108	+ 67.37	+ 131.71	+ 27.42

B.2 Multi-stage Filtering

ClaF uses the multi-stage classification based filtering to find the relevant citations from the whole dataset. The approach to select the classification schemes for each of the three stages is to select classification schemes that are less likely to filter away the true positive citations in the early stages of filtering. Since Naive Bayes (NB) classifier has a higher recall value than that of SVM and 3-NN classifiers, NB classifier is used in the first stage filtering. The 3-NN classifier is used in the second stage filtering, and SVM is used in the final stage of the filtering process.

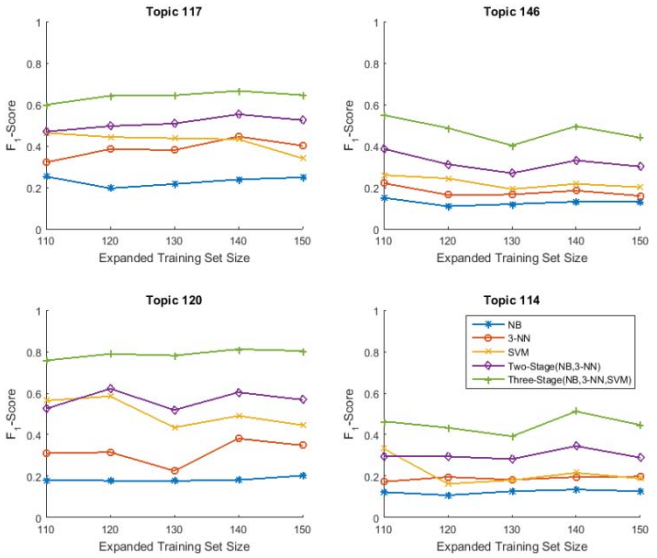


Figure 2: *F_1-Scores* computed for topics 117, 146, 120 and 114 using Three-stage (NB, 3-NN, SVM) filtering, Two Stage (NB, 3-NN) filtering, and NB, SVM and 3-NN only methods.

For comparison purposes, filtering performed with a two-stage process using NB and 3-NN, as well as three one-stage processes with just NB, 3-NN and SVM are also performed. Figure 2 shows the *F_1-Score* for four topics using the five different filtering methods. *F_1-Scores* are calculated using Equation 4. It provides a balanced measure of both recall and precision.

$$F_1 - Score = 2 \cdot \frac{(precision \cdot recall)}{(precision + recall)} \quad (4)$$

As shown in Figure 2, The *F_1-Scores* of the classification results from the three-stage method (NB, 3-NN, SVM) are higher than the other four methods for all four topics.

The final search output is dependent on the order of the text classifiers chosen for the 3 stages. For example a 3-stage filtering with the 3 classifiers in the order: (1) NB, (2) 3-NN, and (3) SVM produces a different result than one with the 3 classifiers in the order: (1) NB, (2) SVM and (3) 3-NN. The first ordering produces a better result with higher *F_1-Scores*. This is because NB is a text classifier that generates classifications with high recall for the given topic. In the second stage 3-NN is used, followed by the SVM text classifier. Since SVM outperformed the other two text classifiers in the one-stage method, SVM is used in the third stage to get an accurate final output.

B.3 Ranked Retrieval Results

ClaF ranks the set of predicted relevant citations from the three-stage process against the *Master Citation* using cosine similarity. The top *N* ranked citations are considered the final retrieval results to be presented to the user for the information need. The retrieval accuracy is computed in terms of the percentage of the top *N* citations having the label of Definitely Relevant (DR) or Possibly Relevant (PR) to the given information need. Table 3 shows the retrieval accuracy of the top 10 citations (P10) and the top 100 citations (P100) in the final search output for the 10 topics.

Table 3: Retrieval accuracy of the top 10 citations (P10) and the top 100 citations (P100) in the final retrieval results.

Topic ID	P10	P100	Topic ID	P10	P100
117	0.9100	0.8462	**109**	0.9520	0.8910
146	0.9100	0.8430	**142**	0.5520	0.6166
120	0.8760	0.7836	**111**	0.6800	0.6698
114	0.8080	0.5740	**107**	0.6700	0.5554
126	0.5820	0.4244	**108**	0.6720	0.3890

It is observed that P10 measure is greater than 0.8 for five information needs. That is, 8 out of the top 10 retrieved citations are relevant to the information need. P10 measure for all the other topics is also greater than 0.55. P100 measure is greater than 60% for six information needs. That is, 60 or more citations from the top 100 retrieved citations are relevant to the information need. Considering that the percentage of citations relevant to each topic is rather small in the entire database, these results are quite encouraging. However, a much lower accuracy is observed for a few topics, e.g., P100 for topic 126 and topic 108. This may be attributed to the fact that there are too few positive citations for the topic. For example, topic 108 has a total of 203 positive citations. For each experiment *n* (*n={5,10,15,20,25}*) positive citations are selected to form the *seed* citation set. The number of remaining positive citations is very small compared to the size of the TREC dataset. This makes the retrieval tasks harder if only the top 100 citations are to be returned. However the P10 and P100 results from the ClaF system present a 13% and 22% improvement over the results reported by the systems during the TREC conference [15]. These results make the ClaF approach a more feasible for personalized retrieval.

V. CONCLUSIONS

The main goal of this study is to build a PIR system based on a small set of input citations. Also, this PIR system focused on retrieving a small set of citations as the search output by eliminating the false-relevant citations. One of the main problems with PIR system is to try to achieve high retrieval quality by training a PIR system using a small set of user provided *seed* citations. In the proposed ClaF, first, the training set is expanded to a reasonably large dataset based on the *seed* citations. Similar citations to the *Master Citation* from the PMRA based dataset are used in expanding the training dataset. This expanded training data allow the NB, kNN and SVM text classification schemes to produce better quality classifiers. Experimental results show that the procedure of expanding training set is successful in achieving its goal. Text classifiers trained from the expanded training set are used in the three-stage filtering method. Three-Stage filtering method is used to successively removing the false positively classified citations from the results. For all the information needs, the *F_1-Scores* of the three-stage method improved dramatically over the base text

classifiers. Also, there is a significant improvement in P10 and P100 measures for a majority of the information needs. Therefore, one can conclude that three-stage filtering method improves the quality of the final search output.

Three-Stage filtering approach can be used for other information retrieval scenarios. The three-stage method may be generalized into multi-stage filtering approach. Our planned next step is to adapt and test the multi-stage method for other domains. We also plan to experiment with using other classification schemes to build the base classifiers. In addition, we plan to experiment with incorporating other feature selection methods and advanced similarity measures into the ClaF system.

REFERENCES

[1] *Google Scholar*. Retrieved 11 16, 2014, from Google Scholar http://scholar.google.com/

[2] *PubMed*. Retrieved 11 16, 2014, from PubMed: http://www.ncbi.nlm.nih.gov/pubmed

[3] The SAO/NASA Astrophysics Data System. Retrieved 11 16, 2014, from Astrophysics Data System: http://adswww.harvard.edu/

[4] *CiteSeerX*. Retrieved 11 16, 2014, from CiteSeerX: http://citeseer.ist.psu.edu/index

[5] *PubMed Fact Sheet*. Retrieved 11 16, 2014, from U. S National Library of Medicine : http://www.nlm.nih.gov/pubs/factsheets/pubmed.html

[6] Islamaj Dogan R, Murray GC, Neveol A, et al. Understanding PubMed user search behavior through log analysis. Database. 2009 doi:10.1093/database/bap018

[7] *PubMed Advanced Search Builder*. Retrieved 11 16, 2014, from U. S National Library of Medicine : http://www.ncbi.nlm.nih.gov/pubmed/advanced

[8] Google Advanced Search. https://www.google.com/advanced_search?hl=en

[9] Mounir Errami, Jonathan D. Wren, Justin M. Hicks, Harold R. Garner: eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Research 35(Web-Server-Issue): 12-15 (2007)

[10] Fontaine JF, Barbosa-Silva A, Schaefer M, et al. MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res. 2009;37:W141–W146.

[11] Poulter G, Poulter G, Rubin D, et al. MScanner: a classifier for retrieving Medline citations. BMC Bioinformatics. 2008;9:108.

[12] Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (2005) 449–456

[13] Qiu F. and Cho J. Automatic identification of user interest for personalized search. In Proc. 15th Int. World Wide Web Conference, 2006, pp. 727–736.

[14] Shen X., Tan B., and Zhai C. Implicit user modeling for personalized search. In Proc. Int. Conf. on Information and Knowledge Management, 2005, pp. 824–831.

[15] Hersh WR, Cohen AM, et al. The Fourteenth Text Retrieval Conference (TREC 2005) NIST; 2005. TREC 2005 Genomics track overview.

[16] Lang, K. Newsweeder: Learning to filter netnews. In Machine Learning: Proceedings of the Twelfth International Conference (ICML '95), pp. 331–339.

[17] Lewis DD, Yang Y, Rose TG, Li F. RCV1: A new benchmark collection for text categorization research. J Mach Learn Res 2004;5:361–97.

[18] Lin J, Wilbur WJ. Pubmed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 2007;8:423

[19] Lee, M.D., Pincombe, B.M., & Welsh, M.B. (2005). An empirical evaluation of models of text document similarity. Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 1254-1259. Mahwah, NJ: Erlbaum.

[20] PubMed Online Training : Related Citations. Retrieved 11 16, 2014, from U. S National Library of Medicine : http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_190.html

[21] Lin J, DiCuccio M, Grigoryan V, Wilbur WJ: Exploring the Effectiveness of Related Article Search in PubMed. In Tech. Rep. LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10. University of Maryland, College Park,Maryland; 2007.

[22] Rorvig M. Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science, Volume 50 Issue 8. 1999.

[23] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schtze. Introduction to Information Retrieval. Cambridge University Press, 2008 pg 234-250,293-320.

[24] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schtze. Introduction to Information Retrieval. Cambridge University Press, 2008 pg 269-277.

[25] Vladimir Cherkassky, Filip M. Mulier. *Learning from Data: Concepts, Theory, and Methods* . WILEY-INTERSCIENCE, 2007.

[26] Burges, Christopher J.C. "A Tutorial on Support Vector Machines for Pattern." Data Mining and Knowledge Discovery, 1998.

[27] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[28] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten; The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (2009).

[29] PubMed Stopwords (11 24, 2014, date last accessed): http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/

Computing the semantic distance between terms: An Ontology-based approach

Alicia Martinez[1], **Fernando Pech**[1], **Noe Castro**[1], **Dante Mujica**[1], **Hugo Estrada**[2], **and Ilse Caspeta**[1]
[1]Computer Science Department, CENIDET, Cuernavaca, Morelos, Mexico
{amartinez, fpech, ncastro, dantemv, ilselanda11c}@cenidet.edu.mx
[2]Research Area, INFOTEC, Mexico City, Mexico
hugo.estrada@infotec.com.mx

Abstract— *The semantic measure determines how they relate two terms or concepts. The challenge of calculating the similarity between terms has become a research area important and has many application in several fields such as artificial intelligence. The development of efficient measures for the computation of semantic similarity is fundamental for computational semantics. Semantic distance is a measure that identifies the strength of relationship between two concepts in an ontology.*

This paper presents the development of novel method (called NaoBig) that expresses the semantic distance between concepts of a knowledge base based on ontologies through a numerical factor. The semantic distance between concepts is shown graphically by a directed graph. Also, BigData RDF is used as search engines and indexing triplets.

Keywords: Semantic distance, semantic similarity measure, path based measure.

1. Introduction

The aim of the Semantic Web is to help automate tasks that require a level of conceptual understanding of the objects involved, and enabling software programs to automatically find and combine information and resources in consistent ways. The core of these new technologies are ontologies [10], which are key to represent formal knowledge so that it can be understood, used and shared between distributed application components.
Ontology is a description (formal knowledge) of concepts and their relationships. However, the information represented in ontology is not always reliable, because there may be two concepts in the same ontology, which are taxonomically distant. For example, given two concepts "heart"and "blood" where "heart" is a subclass of "cardiovascular system" and "blood" is subclasses of "body fluids". Both concepts have no direct relationship within ontology. However, a person might consider that relationship concepts "heart" and "blood" is strong and should have a direct relationship within the ontology under the assumption that the heart pumps blood. To solve this problem, we propose to visualize the ontology to measure the semantic distance between concepts that the user wants to know.

The semantic measures are explored in various fields of research and has various direct and relevant applications such as natural language processing (disambiguation of words [14], synonym detection [13], automatic spelling error detection and correction [3]), knowledge management (thesauri generation [6], information extraction [2], semantic annotation [20], biomedical domain [19], ontologies [7], learning [18], etc.), information retrieval, etc.

The purpose of this paper is to present the development of a novel method (called NaoBig) that expresses the semantic distance between concepts of a knowledge base, which is based on ontologies through a numerical factor. The semantic distance between concepts is shown graphically by a directed graph. Also, BigData RDF is used as search engines and indexing triplets.

The rest of the paper is structured as follows: Section 2 presents related works.. Section 3 describes the method proposed for obtaining the Semantic distance, while Section 4 presents the results of the evaluation conducted during the case studies. Finally Section 5 concludes and briefly discusses future work.

2. Related work

2.1. Semantic measures

The three major semantic measures considered in the literature are: semantic similarity, semantic relation and semantic distance [5]. The semantic similarity is defined taking into account the lexical relations of synonymy (e.g., <car> and <automobile>) and hiperonimia, this measure evaluates the similarity between two concepts of a major subset of semantic links (e.g., is-a and part-of). The semantic relation indicates how distant semantically are two concepts in a network or taxonomy, by using all relations between them (e.g., hyponym, antonyms, meronymy or any functional relation including is-made-of, is-an-attribute-of). The semantic distance is a measure that identifies the strength of the relation between two concepts or terms. If the measure of semantic distance is less, there will be more semantic relationship between the two terms. The similarity measure can be classified by the ontological structure and the content of the information as follows:

- *Path length based measure*. It is based on the distance of the route that separates the concepts or terms. The quantification of similarity is based on the ontology or taxonomic structure [15], [11], [12].
- *Depth relative measure*. It is based on the shortest path approach, considering the depth of the edges of the two concepts in the general structure of the ontology [21], [8].
- *Information content based measure*. It uses both the path length and the depth to determine the similarity between concepts [16].
- *Hybrid measure*. It combines the knowledge derived from several sources of information (such as the path length, local density and some other approaches).
- *Feature based Measure*. It exploits the properties of the ontology to obtain the similarity values and is based on the assumption that each concept is described by a set of words that indicate their properties or characteristics.

In the context of these semantic measures some important contributions have been done.

In [15], Rada proposed an intuitive way to calculate the semantic similarity (also known as taxonomic or attributional measures, it states that the ontologies can be seen as direct graphs in which the concepts are interrelated among them. To calculate the similarity between two nodes/concepts is necessary to count the number of edges in the shortest path between two nodes. This means that the semantic distance of two concepts are correlated with the length of the shortest path.

Given a path $(c_1, c_2) = l_1, ...l_k$ as a set of links that connect the concepts c_1 and c_2 in a taxonomy, and cosidering all possible paths from c_1 to c_2, the semantic distance could be expressed by:

$$dist_{rad}(c_1, c_2) = len(c_1, c_2)$$

where len is the length of the shortest path between c_1 y c_2 with respect to the number of edges. However, this measure is based on the assumption that each edge carries the same amount of information, which does not apply in most ontologies [16].

In [12], Hirst and St-Onge defined the similarity as a distance between the path of two concepts, expressed as follows:

$$sim_{HS}(c_1, c_2) = C - path\ length - k \text{ x } d$$

where d is the number of changes of direction in the path, C and k are constant parameters (the authors use C=8, K=1); if there is no path, $sim_{HS}(c_1, c_2)$ is zero and concepts are not related. Hirst and St-Onge considered the following address path: up (as hiperonimia and meronymy), down (as hyponymy and holonym) and horizontal (as antonymy).

In [8], Ge and Qiu defined the mapping of the similarity of terms based on semantic distance considering the hierarchical relations, relations of semantic distance between terms and degree of measurement mapping between terms through semantic similarity. The algorithm takes two concepts as input and calculates the similarity in four stages: assigning weights between relations, generation of routes or paths between nodes, semantic distance calculation and calculation of semantic similarity. So, given two concepts c_1 and c_2 the expression to compute the weights is next.

$$w[sub(c_1, c_2)] = 1 + \frac{1}{k^{depth}(c_2)}$$

where $depth(c)$ represents the depth of the concept c regard to the concept of the root of node C in the ontology, k is a predefined factor greater than 1 that indicates the rate values of the hierarchy of the ontology.

Wu and Palmer in [21] proposed a strategy to measure the semantic representation of verbs and analyzes the impact on the problems of lexical selection in automatic translations. Since the concepts c_1 and c_2 the similarity measure is calculated with the following expression:

$$Sim_{WP}(c_1, c_2) = \frac{2H}{N_1 + N_2 + 2H}$$

where N_1 and N_2 are the number of "is-a" links from c_1 and c_2, respectively to the Lowest Common Subsumer (LCS) c, and H to the number of "is-a" links from c to the root of the taxonomy.

In [11], Hao et al. used the semantic distance between two concepts by calculating the length of the shortest path c_1 and c_2, as well as the depth under LCS in the tree of lexical hierarchy based on Wordnet to represent different points and calculate the semantic similarity of terms. They propose the following equation to calculate the similarity between two terms:

$$Sim(c_1, c_2) = \left(1 - \frac{\mid path(c_1, c_2) \mid}{\mid path(c_1, c_2) \mid + Depth(LCS(c_1, c_2)) + \beta}\right) \times \left(\frac{Depth(LCS(c_1, c_2))}{\mid path(c_1, c_2) \mid + Depth(LCS(c_1, c_2))/2 + \alpha}\right)$$

where α and β are smoothing factors. When $Depth(LCS(c_1, c_2)) = 0$, both terms have attributes less common and their similarity is 0.

3. NaoBig: Semantic Distance among terms in an Ontology

The semantic distance is a measure that identifies the strength of the relationship between two concepts or terms. You may disagree with the structure of ontology indicating that two terms that are far in the ontology should have a direct relationship. In Figure 1a an ontology that has the words "Aspirin" and "Child" is displayed and can be seen

that there is no relationship between the words; however a user can say that there is a close relationship because a "Child" can take an "Aspirin".

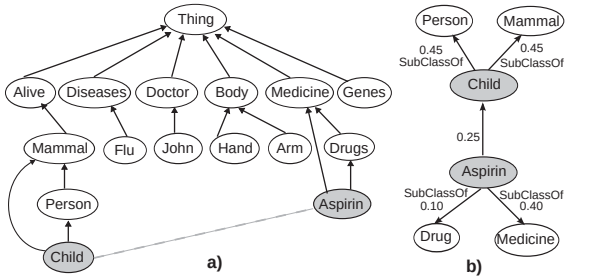


Fig. 1: Partial view of the terms or concepts "child" and "Aspirin"a) in an ontology, b) in an ontology shown by NaoBig.

In this paper a methodology called NaoBig is proposed for the visualization and navigation of a graph showing the semantic distance between concepts from ontology terms entered by the user. Presenting the information in this way, would allow a better decision making. For indexing ontologies and consulting information into NaoBig a BigData RDF API was implemented. Figure 1b shows a partial view of the concepts "Child" and "Aspirin"generated by our method NaoBig, where it can be seen the semantic distance existing between among the concepts.

Figure 2 shows the NaoBig methodology architecture which is made up of three processes: 1) terms and relationship extraction, 2) calculation of semantic distance and 3) generation of graphical and textual information.

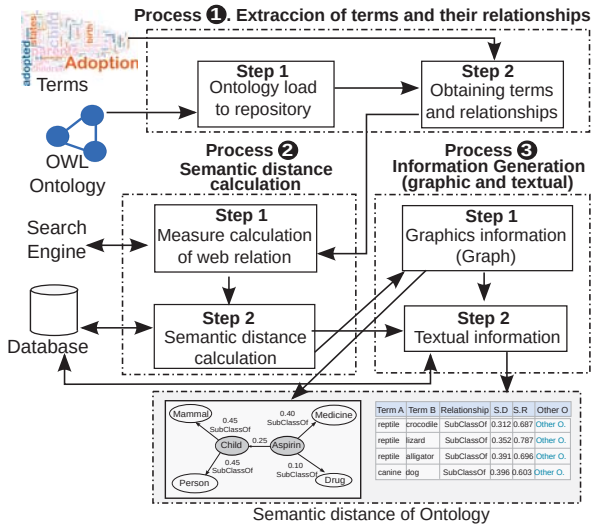


Fig. 2: NaoBig architecture.

3.1. Terms extraction and relationships

This process allows the extraction of all data that can be obtained from an ontology in order to show to the end user those terms related to his query. The process is carried out with two inputs, the terms (to be used in knowledge representation) and an ontology (in OWL). Pairs of terms are extracted from the ontology (they can be superclasses, subclasses, instances or term properties) and type of the relations. The process 1 is made up of two stages:

1. *Loading the ontology in the repository*. OWL Ontology is loaded and saved in a repository in order to access the loaded information trough BigData RDF Database, because of it contains the information that will be used for displaying the semantic distance of terms.
2. *Getting terms and relations*. This stage consists in the retrieving data from the repository containing the ontology; superclasses, subclasses, instances, relations and / or properties that have the terms that the user enters the system are extracted. It can be seen in Figure 1 that the term "child" is related to the terms "person" and "mammal" (both as superclasses). When the user enters any term queries are executed for obtaining the following terms:
   - Retrieving terms of lower level (instances or subclasses).
     $SELECT\ ?z\ ?y\ WHERE\ \{\ ?z\ ?y\ <\ "+term+"\ >\}$
     where *term* is entered by the user. The query returns all those terms $?z$ and the kind of relationship $?y$ that exist with *term*.
   - High level term extraction if the term is an instance
     $SELECT\ ?y\ ?z\ WHERE\ \{<"+term+"\ >\ ?y\ ?z\ \}$
     term may be an instance in the ontology. The query returns terms that can be classes or instances (?z) and the type of relationship (?y) that exists with term.
   - Superclass terms extraction
     $SELECT\ ?z\ WHERE\ \{<"+term+"\ >\ rdfs:subClassOf\ ?z\ \}$

   The results of previous queries are saved in an array that will be used in the process 2.

3.2. Calculating the semantic distance

In this stage a numerical value indicating the degree of relationship (semantic distance) among the terms according to the paternship by frequency is calculated. The pairs of terms are extracted in the process 1. This process is necessary in various APIs that provide data to calculate the semantic distance such as Google Custom Search[1] and Watson[2].

For this process it is necessary a connection to our Database "NaoBig" and Google Custom Search API to

[1]https://developers.google.com/custom-search/
[2]https://developer.ibm.com/watson/

obtain data that help us to calculate the semantic distance. The inputs of this process are the terms and relationships extracted from Process 1 using values obtained with Google Custom Search and Watson. The outputs are the pairs of terms and the semantic distance between terms and their relationships. This process is made up of two stages:

1. *Calculation of Web relationship measure*. The Web relationship among terms is calculated using the Garcia and Mena's formula [9]. The two-term Web relation (from the process 1) is calculated with the measure and frequency of both terms using the formula of the normalized distance of Google $NGD(x, y)$, also known as Normalized Web Distance $NWD(x, y)$ defined in [5] and which is shown above:

$$NGD(x,y) = \frac{max\{log\ f(x), log\ f(y)\} - log\ f(x,y)}{log\ M - max\{log\ f(x),\ log\ f(y)\}}$$

where M is the total number of pages indexed by google or the number of ontologies and semantic Web documents retrieved by watson [17], $f(x)$ and $f(y)$ are the frequencies of terms x and y obtained with the Google Custom Search API and Watson, respectively. Note that Google Custom Search only allows 100 queries per day for free. All frequencies of the query terms in both servers are stored in our database NaoBig so that when you want to know the frequency of a term, it is first checked in our DB; if not found, Google Custom Search and Watson are used.

NGD has a range of values from 0 to ∞, however Gracia and Mena [9] make use of an improvement to this formula to obtain a range from 0 to 1. To obtain the Web relationship next formula is applied:

$$relWeb(x,y) = e^{-2NWD(x,y)}$$

Ontological context (*OC(t)*) is also taking into account, which is a set of ontological terms extracted from the repository containing the ontology, in order to disambiguate the query terms:
OC(t) is defined as the minimum set of ontological terms located on an ontology.

- If *t* is a class then *OC(t)* is the set of direct hyperonyms and is obtained with the following query:
  $SELECT\ ?z\ WHERE\{<\ "+term+"\ > rdfs:subClassOf\ ?z\}$
- If *t* is a instance then *OC(t)* is the class to which it belongs and it is returned with the query:
  $SELECT\ ?z\ WHERE\{<\ "+term+"\ > rdf:type\ ?z\}$
- If *t* is a property then *OC(t)* is the set of classes of its domain and is obtained with the query:
  $SELECT\ ?z\ WHERE\{<\ "+term+"\ > rdfs:domain\ ?z\}$

To calculate the Web relationship of the ontological context the following formula is used:
$relWebOC(x,y) = e^{-2NWD(OC(x),OC(y))}$
where *OC(x)* and *OC(y)* are the ontological context of the term1 and term2, respectively.
The results obtained from *relWeb(x,y)* and *relWebOC(OC(x), OC(y))* are used for calculating the semantic distance.

2. *Calculating the semantic distance*. In this stage the Semantic distance is calculated using the Web relationship calculated in the previous stage. To achieve this, it is necessary to first calculate the semantic relationship between $Term1(x)$ and $Term2(y)$ by applying a weighting w_0 and w_1 to *relWeb(x,y)* and to *relWebOC(OC(x), OC(y))* of the obtained values in Google and Watson.
To calculate the Semantic distance with Google the formula is:
$RSGoogle(x,y) = w_0 * relWeb(x,y) + w_1 * relWebOC(OC(x),\ OC(y))$
Semantic distance in Watson is calculated with:
$RSWatson(x,y) = w_0 * relWeb(x,y) + w_1 * relWebOC(OC(x),\ OC(y))$

where w_0 and w_1[3] are weighting values that must be higher than 0 and the sum must be equal to 1. After obtaining the values of the semantic relationship using Google and Watson, a combination of these two results is performed to obtain the semantic relationship of *x* and *y*, which a weighting to each of the results is again applied, as shown below:

$$RelSem(x,y) = wt_0 * RSgoogle(x,y) + wt_1 * RSwatson(x,y)$$

Where wt_0 y wt_1[4] are weighting values which must be higher than 0 y and the sum of the values must be equal to 1.

According to [9], "the semantic distance is the inverse of the semantic relationship. The two terms more related semantically are the closest to each other". Being 1 the largest value in the semantic relationship and 0 the closest value in the semantic distance, so:
If *Relsem(x,y)=1* then *DistSem(x,y)=0*
If *Relsem(x,y)=0* then *DistSem(x,y)=1*

Therefore, the formula considered for obtaining the semantic distance is as follows:

$$DistSem(x,y) = 1 - RelSem(x,y)$$

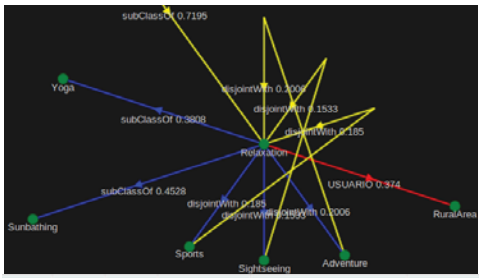
where x and y are the pair of terms obtained from the Process 1. The result of this step is the calculation

[3]The value of w_0 y w_1 applied in this work is 0. 5
[4]The values used in this work are: to $wt_0 = 0{,}7$ and $wt_1 = 0{,}3$

of the distance semantics. This numerical value is the input to the next process that is described below.

3.3. Process Information generation

NaoBig interface for graphical and textual display (see Figure 3) of semantic distance was developed in the third process of our methodology proposed. The JavaScript InfoVis Toolkit[5] was used because provides tools for creating Interactive Data Visualizations for the Web.



Término A	Término B	Relación	Distancia Semántica	Relación Reconocida en la Web (%)	Otras Ontologías
Relaxation	Sightseeing	disjointWith	0.1533	84.67 %	Others Ontologies
Relaxation	Sports	disjointWith	0.185	81.5 %	Others Ontologies
Relaxation	Adventure	disjointWith	0.2006	79.94 %	Others Ontologies
Relaxation	RuralArea	USUARIO	0.374	62.6 %	Others Ontologies
Relaxation	Yoga	subClassOf	0.3808	61.92 %	Others Ontologies
Relaxation	Sunbathing	subClassOf	0.4528	54.72 %	Others Ontologies
Activity	Relaxation	subClassOf	0.7195	28.05 %	Others Ontologies

Fig. 3: Textual and graphical representation of NaoBig.

This process requires the JSON data exchange file (it generated in the process 1) and textual information generated in the process 2. This process consists of two steps:

1. *Graphics information (Graph)*. The JSON data exchange file is generated to graphically display the user's query. It receives as input pairs of terms, relationships and semantics distance calculated in the process 2. The output is both graphic and textual representation of the semantic distance. It also generates a data exchange file that contains the nodes and arrows that compose the graphical representation.
   The data exchange file is composed of three objects, these are: main nodes (these are generated from the terms obtained from the process 2), auxiliary nodes (these are generated from the semantic distance -See Table 1) and relations (these are generated for joining the main nodes and the auxiliary nodes). Infovis JavaScript Toolkit displays objects as Figure 4a.
   For example, if the user creates the relationship between the terms: "Reptile" and "Dog" with a semantic distance of 0.45. This can be seen as in Figure 4b. Thus, 21 auxiliary nodes will be created, as shown in Table 1. Figure 4c shows the terms "Reptile" and "Lizard", with the "SubClassOf" relationship, with a semantic distance of 0.35. therefore auxiliary nodes 17 are created, as shown in Table 1.

[5]http://thejit.org

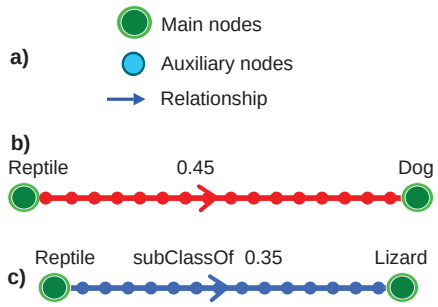


Fig. 4: Representation of the terms on JavaScript Infovis Toolkit.

The output of this step is a graph, which displays the terms related to "reptile" and "dog" and its semantic distance.

Table 1: Number of auxiliary nodes by semantic distance.

S.D.	#nodes	S.D.	#nodes
0.01	3	0.325	16
0.025	4	0.35	17
0.05	5	0.375	18
0.075	6	0.4	19
0.1	7	0.425	20
0.125	8	0.45	21
0.15	9	0.475	22
0.175	10	0.5	23
0.2	11	0.525	24
0.225	12	0.55	25
0.25	13	0.575	26
0.275	14	...	...
0.3	15	1	43

2. *Textual information*. This step generates the textual representation of graphic content in order to display a table with the information contained in the graph. The input of this step is: the pairs of terms, relationships and semantics distance calculated in the process 2.

   The generation of graphic representation required to store the values retrieved in the process 2 in a table (pairs of terms, the relationship and semantic distance) in order to have all the information generated from the user's query. The textual information generated is extracted from the local database.

4. Tests and results

The semantic distance obtained with our methodology was evaluated with two different approaches: correlation and efficiency.

4.1. Correlation approach

The results of the semantic distance obtained with NaoBig are compared with a gold standar WordSim353 [4], [1]. The correlation between the values NaoBig and gold standard

WordSim353 is measured with Spearman correlation (ρ) [1], the following formula is used:

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

The interpretation scale of the correlation coefficient is between -1 and 1, the value 0 indicates no linear association between the two variables of study.

4.2. Efficiency approach

It is for the evaluation of the performance of NaoBig. Pairs of words in different ontologies were taken, a range of values was determined to determining whether two words were closely related, and another range for words with little relationship, the ranges are:

- If the semantic distance >0 and <= 0.5, then, are closely related words (high relationship).
- If the semantic distance >a 0.5 y <1, then, have little related words (low relationship).

The criteria for these measures is based on a heuristic. Since the measure of semantic distance is subjective, we chose to have two ranges; a pair words with high relationship and a pair of words with low relationship, both in the same size range, in this case 0.5. To measure the efficiency of NaoBig, precision metrics, recall and F-measure (efficiency) were used [4].

- To measure the efficiency of NaoBig regarding words with high relationship (*Frel*), the following expression was used:
$$Frel = \frac{2 \times Prel \times Rrel}{Prel \ + \ Rrel}$$
Where (*Prel*) is the precision of words with high relationship, *Rrel* is coverage of words with high relationship.
- To measure the efficiency of NaoBig regarding words with low relationship (*FrelB*), the following expression was used:
$$FrelB = \frac{2 \times PrelB \times RrelB}{PrelB \ + \ RrelB}$$
where (*PrelB*) is the precision of words with low relationship, *RrelB* is coverage of words with low relationship.

The tests were conducted with three ontologies, OntoSem[6] and OpenCyc[7].

For both approaches, 3 distinct formulas were applied to calculate the semantic distance: 1) using the google search engine, 2) using Watson y 3) using NaoBig (combination of both search engines, giving a weight of 0.7 to google and 0.3 to Watson). For reasons of space only two case studies are described:

1. Using the pair of terms "computer" and "software" with OntoSem. This ontology contains 60 pairs words of WordSim353 gold standard used for correlation and efficiency. Table 2 shows that there is a better correlation using Google search engine formula (close to 1). Table 3 shows results in the words with high relationship, and Table 4 shows results in the words with low relationship; in both tables shows that the best results in accuracy, coverage and efficiency (F-measure) are obtained by google.

Table 2: Result of Spearman correlation with OntoSem and OpenCyc.

	Ontosem			OpenCyc		
	Google	Watson	NaoBig	Google	Watson	NaoBig
Spearman	**0.4371**	-0.0427	0.3700	-0.0473	0	**0.13054**

Table 3: Result of precision, coverage, and efficiency in terms (high relationship) with OntoSem.

	Google	Watson	NaoBig
P.T with S.D. <0.5 classifieds manually	39	39	39
Correct number of P.T with S.D.<0.5	26	13	19
P.T <0.5 returned	33	20	37
Precision (high relationship)	**0.7878**	0.65	0.7030
Coverage (high relationship)	**0.7222**	0.4406	0.5757
F-measure (high relationship)	**0.7748**	0.51948	0.5952

P.T. = Pair of terms S.D.= Semantic distance

Table 4: Result of precision, coverage, and efficiency in terms (low relationship) with OntoSem.

	Google	Watson	NaoBig
P.T with S.D. >0.5 classifieds manually	21	21	21
Correct number of P.T with S.D. >0.5	14	14	13
P.T >0.5 returned	27	40	33
Precision (low relationship)	**0.5185**	0.35	0.3939
Coverage (low relationship)	**0.6666**	6666	0.6190
F-measure (low relationship)	**0.5833**	0.4590	0.4814

P.T. = Pair of terms S.D.= Semantic distance

2. Using the terms pair "planet" and "astronomer" with OpenCyc. This ontology contains 29 pairs of words in the gold standard WordSim353. Table 2 shows that the correlation with NaoBig is the nearest to 1. Table 5 shows results with respect to the set of pairs words with high relationship; therefore, the better accuracy and coverage are of Watson, and F-measure by NaoBig. Table 6 displays the results in the words with low relationship: coverage by Watson, precision and F-measure by NaoBig.

Of the various tests with gold standard, the better correlation was with NaoBig. With an efficiency of almost 60 % regard to the semantic distance, precision of up to 80 % in the pairs of terms with high relationship, and the low relationship terms was obtained less than 70 %.

[6]http://morpheus.cs.umbc.edu/aks1/ontosem.owl
[7]http://www.OpenCyc.com/platform/openOpenCyc/downloads

Table 5: Result of precision, coverage, and efficiency in terms (high relationship) with OpenCyc.

	Google	Watson	NaoBig
P.T with S.D. <0.5 classifieds manually	20	20	20
Correct number of P.T with S.D. <0.5	3	1	8
P.T <0.5 returned	5	1	9
Precision (high relationship)	0.6	**1**	0.8889
Coverage (high relationship)	0.15	0.05	**0.4**
F-measure (high relationship)	0.2400	0.09524	**0.5517**

P.T. = Pair of terms S.D.= Semantic distance

Table 6: Result of precision, coverage, and efficiency in terms (low relationship) with OpenCyc.

	Google	Watson	NaoBig
P.T with S.D. >0.5 classifieds manually	9	9	9
Correct number of P.T with S.D.>0.5	7	9	8
P.T >0.5 returned	24	28	20
Precision (low relationship)	0.29167	0.32143	**0.4**
Coverage (low relationship)	0.7778	**1**	0.88889
F-measure (low relationship)	0.4242	0.48649	**0.5517**

P.T. = Pair of terms S.D.= Semantic distance

5. Discussion and future work

The semantic distance calculation depends on the information that is retrieved by the search engines (google and Watson). This is because the process of semantic distance calculation is based on the relationship of association frequency of terms in the corpus. NaoBig, a methodology based on the combination of values obtained with the search motors Normal Web and Semantic Web was developed in this paper. As we have experimentally demonstrated, the results obtained with our approach were better than those obtained with each search motor evaluated in separately way. Quantitatively, the proposed scheme obtained an efficiency close to 60 % with respect to semantic distance, an accuracy of 80 % to recognize terms with high relation, a coverage of 76 % to recognize terms with low relation. Also, it obtained the best correlation coefficient in comparision with Google and Watson. To extend the tool to future,be advisable to perform different tests with a larger set of pairs of terms from different ontologies. Furthermore, it would be interesting to add another numerical factor to calculate the semantic distance, as the frequency of terms from a corpus.

References

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA, 2009.

[2] John Atkinson, Anita Ferreira, and Elvis Aravena. Discovering implicit intention-level knowledge from natural-language texts. *Knowledge-Based Systems*, 22(7):502 – 508, 2009.

[3] Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pages 29–24, Pittsburgh, PA, 2001.

[4] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32(1):13–47, March 2006.

[5] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, March 2007.

[6] James R. Curran. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 222–229, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[7] Anna Formica. Concept similarity in formal concept analysis: An information content approach. *Knowledge-Based Systems*, 21(1):80 – 87, 2008.

[8] Jike Ge and Yuhui Qiu. Concept similarity matching based on semantic distance. In *Proceedings of the 2008 Fourth International Conference on Semantics, Knowledge and Grid*, SKG '08, pages 380–383, Washington, DC, USA, 2008. IEEE Computer Society.

[9] Jorge Gracia and Eduardo Mena. Web-based measure of semantic relatedness. In *In Proc. of 9th International Conference on Web Information Systems Engineering (WISE 2008), Auckland (New Zealand*, pages 136–150. Springer, 2008.

[10] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.

[11] Dou Hao, Wanli Zuo, Tao Peng, and Fengling He. An approach for calculating semantic similarity between words using wordnet. In *ICDMA*, pages 177–180. IEEE, 2011.

[12] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms, 1997.

[13] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[14] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 241–257, Berlin, Heidelberg, 2003. Springer-Verlag.

[15] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[16] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[17] Marta Sabou, Miriam Fernández, and Enrico Motta. Evaluating semantic relations by exploring ontologies on the semantic web. In *Natural Language Processing and Information Systems, 14th International Conference on Applications of Natural Language to Information Systems, NLDB 2009, Saarbrücken, Germany, June 24-26, 2009. Revised Papers*, pages 269–280, 2009.

[18] David Sánchez. A methodology to learn ontological attributes from the web. *Data & Knowledge Engineering*, 69(6):573 – 597, 2010.

[19] David Sánchez and Montserrat Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44(5):749 – 759, 2011.

[20] David Sánchez, Montserrat Batet, and David Isern. Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297 – 303, 2011.

[21] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

SESSION

INFRASTRUCTURES + APPLICATIONS OF INFORMATION AND KNOWLEDGE ENGINEERING + LEGISLATION AND INTELLECTUAL PROPERTY DISCUSSIONS

Chair(s)

TBA

Wind Power Generation Prediction on a Large Real Life Dataset

Roshil Paudyal[1], Mugizi Robert Rwebangira[2], Mandoye Ndoye[3], roshil.paudyal@bison.howard.edu, rweba@scs.howard.edu, mndoye@mytu.tuskegee.edu

[1]Department of Mathematics, Howard University, Washington, D.C. 20059 USA
[2]Systems and Computer Science, Howard University, Washington, D.C. 20059 USA
[3]Department of Electrical Engineering, Tuskegee University, Tuskegee, AL 36088

***Abstract*— The intermittent nature of wind power generation is becoming more of a problem as the percentage of wind energy used in the grid is increasing. We propose a data driven approach using machine learning methods to predict daily wind power generation output. A novel aspect of this work is the verification of the algorithm on a massive dataset with more than 500,000 observations.**

Keywords- Wind Power; Renewable energy; Machine Learning ; Logistic regression

I. INTRODUCTION

Wind power accounts for a relatively small percentage of energy in the US distribution grid despite the fact that it constitutes one of the most promising sources of clean and renewable energy. When the percentage of wind energy in the grid is negligible, control room operators can schedule these resources without facing any serious issues. However, the percentage of wind energy usage within the grid has been steadily increasing, and there is now a pressing need for more accurate forecasts of wind power production, which can then be exploited to make better informed scheduling decisions. The integration of wind energy in the power grid is a very difficult task because of its intermittent nature. One of the main challenges is the lack of predictability of the amount of power from wind turbines. This increases the spinning reserve requirements and unanticipated ramp events, causing elevated production costs and decreased reliability. Accurate and reliable methods for forecasting wind power generation are essential if wind power is to become a staple in countries energy diet. In this work we address this problem. One of our main contribution is that we are able to verify our algorithms on a dataset with 500,000 observations, which is far larger than most previous work.

II. RELATED WORK

Several researchers have proposed algorithms for wind power forecasting. Just to mention a few of them, Mabel and Fernandez [6] proposed using artificial neural networks [ANN] for wind power prediction. Their data covers only a 2 year time span. Świątek and Dutka [4] also propose a neural network approach. Their data covers less than a 2 year time period. Finally Lei et. al [7] also give a comprehensive bibliography of various approaches. Many of them were preliminary and only used limited data or were test runs.

III. THE DATA

Our data, retrieved from the Bonneville Power Administration [4] has one of the largest observations of weather and power output available. The data records Barometric Pressure, Humidity, Temperature, Wind Speed, Wind Direction, Peak Wind Speed and Peak Wind Direction at the station at each 5 minute interval from 01/01/2009 to 01/01/2014 or a period of more than 5 years.

Below is an example of what the data looks like:

Date/Time (UTC)	Barometric Pressure (INHG)	Relative Humidity (PT)	Temperature (F)	Wind Direction (DEG)
2/1/10 8:00	26.7	89.6	37	44.9
2/1/10 8:05	26.7	90.4	37.2	44.9
2/1/10 8:10	26.7	91.1	37.6	44.9
2/1/10 8:15	26.7	91.8	37.6	44.9
2/1/10 8:20	26.7	91.3	37.7	51.8

We also have the corresponding power output at each time period.

IV. ALGORITHM

As a preliminary step, we are just trying to predict the wind power output as high or low value using a logistic regression model. In order to train the model, we convert a continuous-valued power output variable to a binary one. We use the median of the power output as a criteria to judge the power output values as high or low, i.e. a value higher than the median (1063 MW) was given the binary value 1, and a value lower than that was given the binary value 0.

We then used Pressure, Humidity, Temperature and Wind Speed as predictors and trained the logistic model to output a binary high/low value given those parameters. One of the issues with the data was that while it contained individual files describing the weather at individual stations, the power output provided was the sum of power output by all stations, which made it challenging to assess the contributions of the individual stations to the overall power output. So we devised a step-by-step approach, which first takes the weather data at a single station as a predictor, and then averages the weather

Figure 1: Output of logistic regression with a randomly chosen station as the predictor

Model:	Logit	Df Residuals:	49996
Method:	MLE	Df Model:	3
Date:	Tue, 31 Mar 2015	Pseudo R-squ.:	0.06909
Time:	20:07:47	Log-Likelihood:	-32227.
converged:	True	LL-Null:	-34618.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Pressure	-0.1235	0.003	-40.367	0.000	-0.130	-0.118
Humidity	0.0072	0.001	11.095	0.000	0.006	0.008
Temperature	0.0334	0.001	37.122	0.000	0.032	0.035
Wind_Speed	0.0851	0.001	59.711	0.000	0.082	0.088

data across all the stations, and takes that as a predictor of high or low power output.

V. RESULTS

In Fig 1 we show the output of logistic regression when we use a randomly chosen station's weather data as the predictor.

Especially interesting to note is the correlation of the various weather indicator's with the power output. As we expect the correlation is relatively small, and even negative for pressure.

Then we look at the raw accuracy of the prediction, again with the weather at one randomly chose station:

	Predicted high	Predicted low	% Correct
Observed high	4890	2707	64.36
Observed low	7727	5224	59.67
Overall			61.40

The overall accuracy of 61% is actually very respectable and compares favorably with previous work.

In Fig. 2 we take the average of the weather data across 14 stations and use those averages as our features.

We immediately note that the correlation is much higher between wind speed and power output, which is what we'd expect.

Looking at the raw accuracy of the predictions:

	Predicted high	Predicted low	% Correct
Observed high	5502	2101	72.36
Observed low	9909	3036	76.54
Overall			75.00

The accuracy has increased from 61% when using only one station to 75%, which is state of the art in comparison with results reported in the literature.

VI. CONCLUSION

We can draw a couple of conclusions from these preliminary experiments.

(1) The massive amount of observations directly lead to more accurate prediction and are effectively utilized (in comparison with other researchers who had access to less data).

(2) Even though taking the average of the weather across different stations is not a perfect solution it still leads to remarkably improved prediction accuracy.

Figure 2: Output of logistic regression with the average of 14 stations as the predictor

Model:	Logit	Df Residuals:	49996
Method:	MLE	Df Model:	3
Date:	Tue, 31 Mar 2015	Pseudo R-squ.:	0.1926
Time:	18:27:24	Log-Likelihood:	-27958.
converged:	True	LL-Null:	-34627.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Pressure	-0.2431	0.005	-52.852	0.000	-0.252	-0.234
Humidity	0.0156	0.001	18.265	0.000	0.014	0.017
Temperature	0.0351	0.001	28.120	0.000	0.033	0.038
Wind_Speed	0.4405	0.005	93.197	0.000	0.431	0.450

VII. FUTURE WORK

A more principled way of combining the weather observations from different stations is highly desirable. One way of doing this might be by taking an average of the weather variables weighted by their correlation with the power output.

REFERENCES

[1] Bonneville Power Authority Load and Total Wind Generation, http://transmission.bpa.gov/business/operations/wind/

[2] Javad Mahmoudi, Majid Jamil, Hossein Balaghi. Short and Mid-Term Wind Power Plants Forecasting With ANN, Second Iranian Conference on Renewable Energy and Distributed Generation, 2012

[3] Atsushi Yona, Tomonobu Senjyu, Funabashi Toshihisa, Chul-Hwan Kim, Very Short-Term Generating Power Forecasting for Wind Power Generators Based on Time Series Analysis, Smart Grid and Renewable Energy, Vol.4 No.2, May 2013

[4] Bogusław Świątek, Mateusz Dutka. Wind power prediction for onshore wind farms using neural networks. International Conference on Renewable Energies and Power Quality. 2015

[5] Lawrence Livermore National Labs, Predicting Wind Power with Great Accuracy, Science and Technology Review, April/May 2014

[6] M. Carolin Mabel, E. Fernandez. Analysis of wind power generation and prediction using ANN: A case study, Renewable Energy Volume 33, Issue 5, May 2008, Pages 986–992

[7] Ma Lei, Luan Shiyan, Jiang Chuanwen, Liu Hongling, Zhang Yan. A review on the forecasting of wind speed and generated power, Renewable and Sustainable Energy Reviews Volume 13, Issue 4, May 2009, Pages 915–920

The impact of uncertainty measures in the cardiac arrhythmia detection supported by IK-DCBRC

Abdeldjalil KHELASSI

Department of informatics , Abou Bakr Belkaid University, Tlemcen, Algeria

Abstract - *The multi-agent systems is a distributed artificial intelligence approach in which a set of agents collaborate to achieve the global goal of the system. This paradigm affects several quality factors of complex systems as performance, transparency and accuracy. This work describes an original empirical experiments realized by IK-DCBRC, which is a multi-agent system for medical decision support. The distributed application is based on a cognitive amalgam, where conflicts and contradictory decisions can be detected and visualized via an explanation agent. In this paper, we introduce a new parameter, fuzzy uncertainty measures, for an appropriate aggregation of agent's decisions. We have realized some empirical experiments by using the cardiac arrhythmia data extracted from physiological signal.*

Keywords: Uncertainty measures; Fuzzy sets theory; Multi-Agent System; Cardiac arrhythmias; Case-Based Reasoning.

1 Introduction

The uncertainty is an important problem concerned by several active research area as statistics [10], communication sciences [12], finance and economic sciences [13] and artificial intelligence [4,5 and 11]. Three type of uncertainty is defined in [4], which is: 1-non-specificity (imprecision) 2- vagueness and 3-strife discord.

To deal with the problem of uncertainty, Lotfy Zadah in 1965 introduced the Fuzzy Sets Theory FST. It gives a new extension of the traditional theory sets [5]. Several uncertainty measures was introduced in the domain of fuzzy sets presented in [4] according to the type of uncertainty.

The distributed reasoning systems via cognitive agents ensure several quality factors for intelligent systems as transparency, accuracy and performance. It is an appropriate paradigm for parallel reasoning and distributed computing [15, 18 and 19].

The case-based reasoning CBR paradigm is successfully applied in several health science applications [15, 16]. It consists to resolve new problems via the solutions of similar cases [14, 15]. The distributed case-based reasoning DCBR is a variant of traditional CBR in which; the reasoning is distributed through cognitive agents and the cases through a set of case bases [15].

In this paper, we will describe the uncertainty measures in fuzzy sets theory, The IK-DCBRC software applied in the detection of cardiac arrhythmias, the empirical experiments and a short discussion.

2 Uncertainty measures

2.1 Fuzzy sets

The fuzzy sets [4, 5] generalize the classical sets by considering the membership as a graded concept. The membership degree of an element x to a fuzzy set A denoted by $\mu_A(x)$, take a value in the interval [0, 1] (see Figure1).

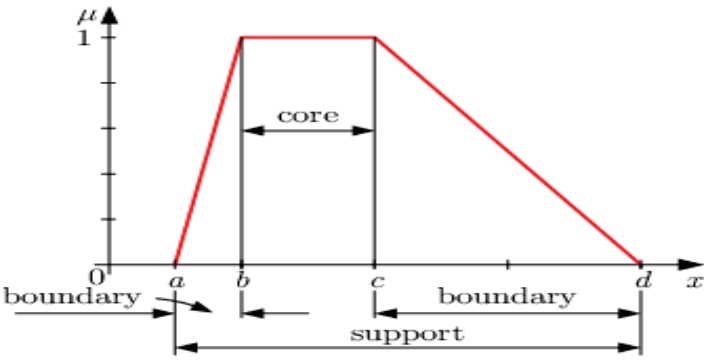


Fig. 1. The memcership function.

The support of a fuzzy set A is the crisp set that contains all the elements of X that have nonzero membership grades in A.

$$supp(A)=\{x\epsilon X,\ \mu_A(x)>0\} \quad (1)$$

The core of a normal fuzzy set A is the crisp set that contains all the elements of X that have the membership grades of one in A.

$$core(A)=\{x\epsilon X,\mu_A(x)=1\} \quad (2)$$

The boundary is the crisp set that contains all the elements of X that have the membership grades of $0< \mu_A(x) <1$ in A.

$$bnd(A)=\{x\epsilon X,0<\mu_A(x)<1\} \quad (3)$$

Having two fuzzy sets A and B based on X, then both are similar if:

Core (A)=Core (B) and Supp(A)=Supp(B) (4)

2.2 Uncertainty Measures

The uncertainty measure is well described in [5].It represents a common problematic between several domains as mathematics, physics, and cognitive science. In fuzzy sets theory [5] the uncertainty measure is defined as:

Definition1: The α-cut of a fuzzy set A is

$$^{\propto}A = \{\propto | \mu_A(x) \geq \propto\} \quad (5)$$

And a strong α-cut

$$^{\propto+}A = \{\propto | \mu_A(x) > \propto\} \quad (6)$$

Definition2: The uncertainty of a fuzzy set A is measured with the function U:

$$U : P(A) = \{\phi\} \rightarrow \mathbb{R}^+$$
$$U(A) = \frac{1}{h(A)} \int_0^{h(A)} \log_2 |^{\propto} A| d\propto \quad (7)$$

Where $|\propto A|$ denotes the cardinality of the α-cut and h(A) the height of A. This function is called also the non-specificity function.

3 Case-based reasoning

3.1 Definition

The case-based reasoning paradigm is successfully applied in several crucial domains as medicine [1, 2, 3, 8 and 15], industry [6, 7 and 14], information retrieval [14], image processing [14] and others. It is an important applied method for classification and clustering [1, 2, 3, 6 and 7].

We have defined the case-based reasoning in [15] as: "An intelligent approach inspirited from the human reasoning. It consists to use the prior expertise to resolve a new problem. This expertise or knowledge is constructed as a set or collection of cases. Each case represents problem associated with its solution. The idea is that two similar problems have the same solution. Then to resolve a new problem we will pass by the similarity measures between this problem and all cases in the case base. The expert can add a new knowledge (adapted cases) then we can considerate the CBR as a machine learning technique".

Global-local similarity measures

The global-local similarity measure is based on decomposing the entire similarity computation in a local part, in which only similarities between single attribute values is considered locally, and a global part computing the global similarity for whole case based on the local similarity assessments. Such decomposition simplifies the modeling of similarity measures significantly and allows defining well-structured measures even for very complex case representations consisting of numerous attributes with different value types. In this section, we will define some concepts of this approach.

Definition3. (The weight vector)

Let D = (A_1, A_2,.., A_n) be a case characterization model. The vector v (w_1, w_2,..,w_n) with w_i in [0, 1] and $\sum w_i=1$,is called weight vector for D, where each element w_i is called attribute weight for A_i.

Definition4. (The local similarity)

A local similarity measure for an attribute A is a function

$Sim_A : A_r \times A_r \rightarrow [0, 1]$.

With A_r is the value range of the attribute A. There are many developed similarity function for the local measures as linear, threshold, exponential, sigmoid, Cosin and other similarity function, which permit the computing of local similarity function between tow attributes with the same domain. Some one used the Euclidian distance other uses the logarithmic distance for the numerical attribute and the similarity table or ontological distance for words. In some works, they combine between two similarity functions one for negative distances and one for positive distances [15, 9].

For example, the sigmoid Similarity function is defined as:

Sim_i: DxD→[0,1]

$$Sim_i(q_i, c_i) = \frac{1}{1 + e^{\frac{\delta(q_i, c_i) - \theta}{\alpha}}} \quad (8)$$

With q_i and c_i are the attributes number i of the query Q and the case C.

D: denotes the space of case characterization models.

The parameters α and θ are defined intuitively after some experiments.

The σ function represents the distance between the attributes; generally, they use for this distance function the Euclidian (10) or logarithmic function (9).

The logarithmic distance function

σ:DxD→IR

$$\sigma(q_i, c_i) = \begin{cases} \ln(c) - \ln(q) & \text{for } q, c > 0 \\ \ln(-c) - \ln(-q) & \text{for } q, c < 0 \\ \text{Undefined} & \text{else} \end{cases} \quad (9)$$

The Euclidian distance function

$$\sigma(q_i, c_i) = |q_i - c_i| \quad (10)$$

Definition5. (The global similarity)

Let D = (A1,A2, . . ., An) be a case characterization model, w be a weight vector, and sim_i be a local similarity measure for the attribute Ai. A global similarity measure for D is a function

Sim : $D_D \times D_D \rightarrow [0, 1]$

$Sim(q, c) = \pi(sim_1(q.a1, c.a1), \ldots, sim_n(q.an, c.an), w)$ (11)

where $\pi : [0, 1]^2 \rightarrow [0, 1]$ is called aggregation function that must fulfil the following properties:

$$\forall \vec{w} : \pi(0,..,0,\vec{w}) = 0$$

π is increasing monotonously in the arguments representing local similarity values.There are many defined aggregation function π in the stat of the art we can cite in this section some examples of commonly used [15].

- **Weighted Average Aggregation**
$$\pi(sim_1,..,sim_n,\vec{w}) = \sum_{i=1}^{n} w.sim_i$$
- **Minkowski Aggregation**
$$\pi(sim_1,..,sim_n,\vec{w}) = (\sum_{i=1}^{n} w.sim_i^p)^{1/p}$$
- **Maximum Aggregation**
$$\pi(sim_1,..,sim_n,\vec{w}) = max_{i=1}^{n} w.sim_i$$
- **Minimum Aggregation**
$$\pi(sim_1,..,sim_n,\vec{w}) = min_{i=1}^{n} w.sim_i$$

3.2 Importance degree of features

The importance degree of features is generally called the attributes weights. The modeling of the values of these weights is not a common standard process in the developed CBR systems. Several approaches used in the developed case-based reasoning systems for characterizing the features weights; we present in this section the following approaches [15]:

1. Global Weights: The developer defines globally the weight model for the application. This is the most general weight model for the definition of the importance of attributes in CBR systems. In this approach, the defined weights are valid for the entire application domain, and by consequence: the influence of attributes on the utility approximation is constant for all cases and queries that may occur.

2. Case Specific Weights: This more fine-grained weight model allows the description of different attribute weights for different cases. This means, when comparing a query with a given case, a specific weight vector for this particular case is used to perform the similarity computation. A special form of this weight model is class specific weights used for classification tasks. Here, the weight vector to be used is determined by the class membership of the particular case.

3. User Weights: Another approach is the use of specific weights for each new retrieval task, i.e. the weights are acquired together with the query. Such a weight model is in particular useful in domains where the users might have individual preferences with respect to the utility of cases. For example, a product recommendation system in e-Commerce might allow customers to input attribute weights in order to express the importance of particular product properties for the transaction.

Different weight models can also be combined. For example, user weights are often not used exclusively, but they are combined with a global weight vector defining the general importance of attributes from the application domain point of view.

4 IK-DCBRC

4.1 Description of the software

We have developed a CBR framework IK-DCBRC for pattern classification, which contains two kinds of cognitive agents: Adaptation agent and Similarity agent (see Fig.2.). Each case base contains cases from the same class. Each agent uses a predefined knowledge, which contains ontology, rules and heuristics to realize their local goals. The main goal of the system is to generate the class of the query, but each agent is autonomic for satisfying its local goals, which are:

1. the degree of membership of the query in the fuzzy sets of the associated class for the similarity agents.
2. editing queries and the adaptation of the retrieved cases for defining the queries class for the adaptation agent.

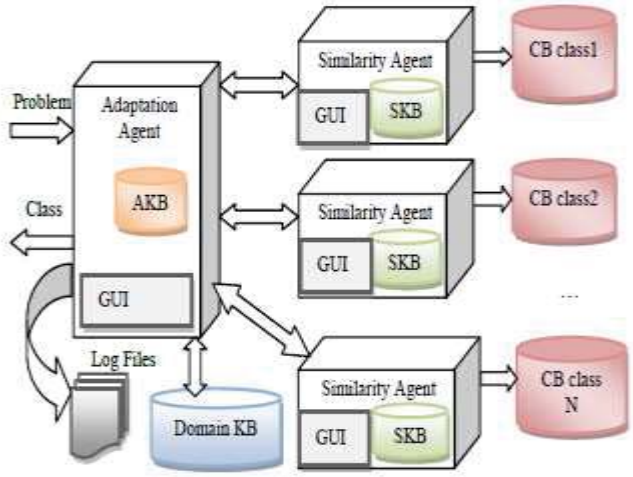


Fig. 2. The IK-DCBRC architecture

Each agent has a specific General User Interface for introducing the data and classification parameters (test base, case bases, weight vectors, similarity parameters and the

adaptation knowledge base) also for applying some optimization algorithms.
The interactive interfaces are used also for the explanation [8] and the retaining processes. These interfaces assure a high level of explanation and flexibility, which we can't find in another framework or classifier.

4.2 Cognitive Agents description

The developed system contains two kinds of cognitive agents, the similarity agents and adaptation agent. The system infers the decision through a collaborative behavior between all the existing agents.

The similarity agent aims to compute the similarity between the query and some cases from the same class. Several similarity measures were defined and can be selected via the Graphical User Interface see fig3.

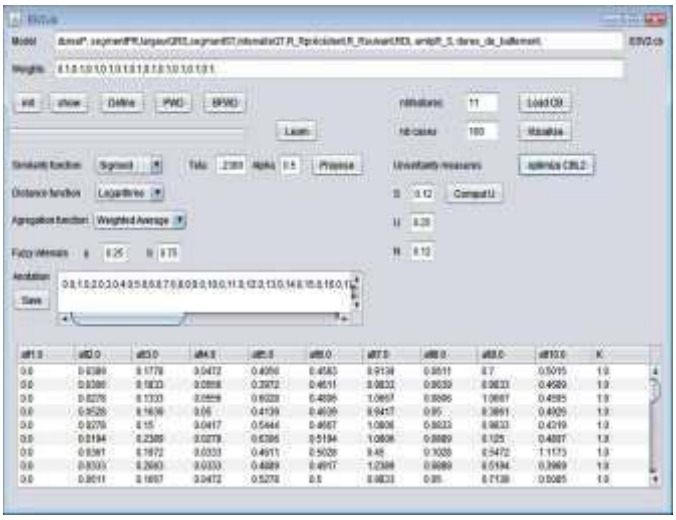


Fig. 3. Similarity agents GUI

The adaptation agent aim is to generate the appropriate decision after collecting the responses of all agents. It infer also from a rule-based system, which contains a partial domain knowledge base.

Each cognitive agent constructs its knowledge via some machine learning algorithms defined in [1].

4.3 Uncertainty measures in IK-DCBRC Applications

An original similarity function is introduced in the IK-DCBRC and evaluated see [1, 2 and 3]. This function is based on three fuzzy sets for the linguistic variables similar, not similar and unknown. The proposed approach is developed for increasing not just the accuracy but for ensuring also the flexibility and the transparency of the reasoning process. It combines the local-global similarity functions and the fuzzy sets theory. With the novel similarity function not just the traditional response is generated, which is one real value between 0 and 1, but it generates three values. Three possible responses can be inferred by the similarity function: 1- The unknown response, which represents the negative response, is inferred when the similarity agents generates a high degree of membership in the unknown set (14). 2- The non-similar response when the similarity agent generates a high degree of membership in the non-similar set (13). 3- The similar response when the similarity agent generates a high degree of membership in the similar set (12).

For accomplishing the criteria we have defined three fuzzy sets similar S, not similar N and unknown U with the membership functions μ_S, μ_N and μ_U. These membership functions are defined as follow:

$$\mu_s(x) = \begin{cases} 0 \ if \ x \le a \\ \frac{x-a}{1-a} \ if \ x > a \end{cases} \quad (12)$$

$$\mu_n(x) = \begin{cases} 0 \ if \ x \ge b \\ \frac{b-x}{b} \ if \ x < b \end{cases} \quad (13)$$

$$\mu_u(x) = \begin{cases} 0 \ if \ x \le b \ ou \ x > a \\ \frac{x-b}{0.5-b} \ if \ x > b \ and \ x \le 0.5 \\ \frac{a-x}{a-0.5} \ if \ x < a \ and \ x > 0.5 \end{cases} \quad (14)$$

In this contribution, we have used the triangular function for the fuzzy sets. The variable x represents the result of the similarity aggregation function between the query and the case. The support of the fuzzy sets is defined intuitively by using the agents GUI or by using a machine-learning algorithm see fig 3. Also the formula of x is selected by the user and it represents the global similarity measures between the query and the case defined in (8).
Each similarity agent measures the uncertainty via the formula (7) (see fig3), and each decision is associated with the uncertainty value according to the reasoning process.

5 Empirical experiments

The aim of this research work is to presents the importance of uncertainty measures in the detection of cardiac arrhythmias. The first use of IK-DCBRC for this application was in 2008, for the detection of cardiac arrhythmia via the characteristics of beats described in table1 [17].
The cardiac beat model in table1 represents the characteristics used in the learning and evaluation steps. This model is defined by using some computational methods explained in [15].

Attribute	Type	Description
Pdur	REAL	The duration of the wave P.
PRseg	REAL	The PR segment.
QRS	REAL	The QRS larger.
STseg	REAL	The ST segment.
QTInterval	REAL	The QT Interval.
R_priorR	REAL	Distance between the current R and the prior one.
R_nextR	REAL	Distance between the current R and the next one.
RDI	REAL	R_priorR/ R_nextR
AmpR_S	REAL	Distance between R and S.
Beat_duration	REAL	The Beat duration.
Old_type	String	The Age kind(Adult/child)
ECG_Leads	String	The ECG lead.

TABLE I. THE DATASET FEATURES

In these experiments, we have used the sigmoid function for each similarity agent. We have also applied the global weights strategy, by considering the features with equal importance in experiment 1 and learned from data in experiment 2 (by introducing the uncertainty measures in the learning step). We have defined the fuzzy sets supports intuitively (a=0.25 and b=0.75) i.e a fixed uncertainty U=0.125 for all experiments.
The adaptation agent infers a partial Domain Knowledge Base DKB coded in XML. The DKB is well described in [1 and 15]. For improving the Adaptation Knowledge Base AKB, we have constructed another data set, from the log files, which record the trace of reasoning of each agent. This data set contains 200 instances and 8 features (uncertainty measures computed by similarity agents, the response inferred from the DKB and the class). The number of features of this data set is according to the number of similarity agents.

6 Results

In figure 6 and 7, we present the results of correct classification and errors with several approaches used for the aggregation of agents decisions. In the described experiments we have used a data set which contains a classified cardiac beats characterized by 11 features presented in table1 [1].

In these experiments, we have used the adaptation rules applied in [1,2 and 3] but the parameters described above (see section V). In addition, we have applied some non-symbolic approaches as Support Vector Machine SVM, Artificial Neuronal Network, Bayesian Network BN and Decision Tree. The WEKA software generates all non-symbolic experiments results.
In experiment1, we have fixed the features weights with the same value i.e 1⁄N=0.1 see figure 4.

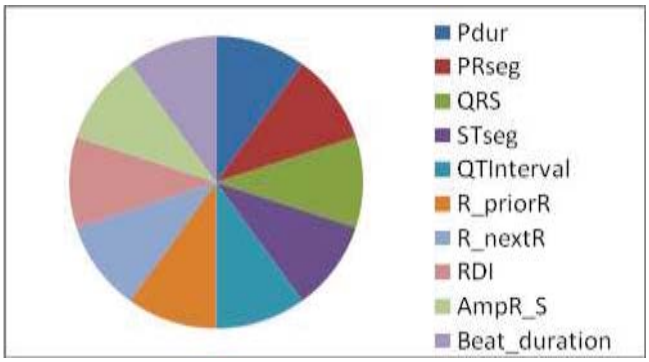


Fig. 4. Experiment 1:The importance degree of features all weights w_i=0.1.

In experiment 2, we have used the weights defined by a machine-learning algorithm, fuzzy gradient decent variant, which is described in [1 and 15]. The obtained weights are visualized in figure 5.

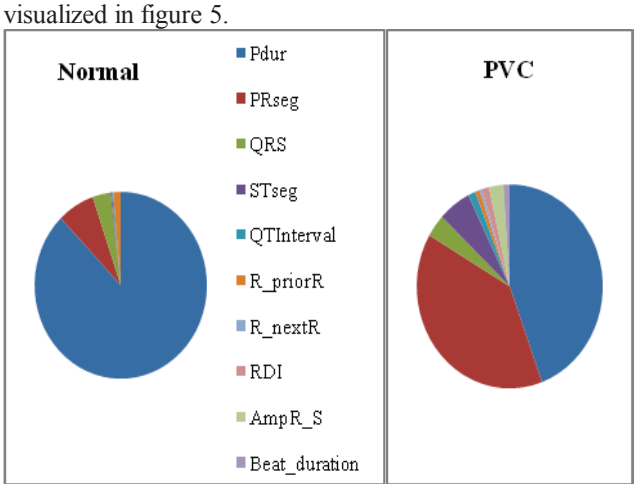


Fig. 5. Experiment 2:The importance degree of features defined by the gradian descent algorithme and by concidiring uncertainty measures.

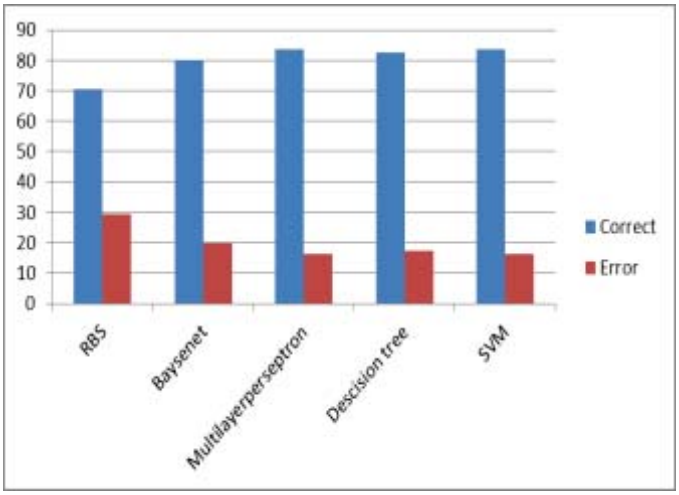


Fig. 6. Experiment1: The results of classification by IK-DCBRC whre the features weights are equal i.e w_i=0.1

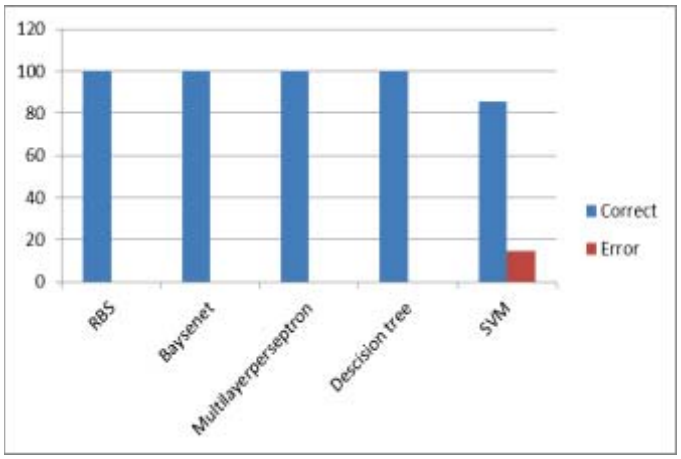


Fig. 7. Experiment2:Results of classification for fuzzy machine learning definition of weights

7 Discussion

In the experiment1 the uncertainty measures is not considered for the weights definition. The obtained rate of correct classification was between 70% and 83,75% and the error between 18,25% and 30%.

In the experiment 2 each agent applies the gradient descent algorithm which measures the error by including the uncertainty measures. The obtained results was optimal 100% for all methods just the SVM, the rate of correct classification was 85,5%.

The impact of uncertainty measures is very clear by comparing the results of experiment1 and experiment2. All computational methods symbolic, non-symbolic or probabilistic improve this impact.

In experiment1, the correct classification is just 70%. Although the rule-base systems are the most transparent intelligent systems, in these experiments the RBS represents inflexibility with the domain uncertainty.

The errors of SVM in experiment 2 is less than it's errors in experiment1, which reinforce the conclusion. These errors are due to the inflexibility with the treated uncertainty in this domain. However, we can conclude the strong side of consideration of uncertainty measures in the definition of features weights.

The non-symbolic methods, i.e decision tree and artificial neuronal network also the Bayesian network, present a good flexibility with uncertainty in experiment 1 and an optimal flexibility in experiment 2.

8 Conclusion

In this paper, we have the opportunity to treat some aspects of uncertainty measures in a successful multi-agent system IK-DCBRC. The impact of uncertainty measures in the detection of cardiac arrhythmia is well described via the realized empirical experiments. The uncertainty measures ensure the transparency of reasoning in joint with the accuracy of detection.

9 References

[1] Khelassi, A. and Chikh, M.A. (2015) 'Cognitive amalgam with a distributed fuzzy case-basedreasoning system for an accurate cardiac arrhythmias diagnosis', Int. J.Information and Communication Technology, Vol. 7, Nos. 4/5, pp.348–365.

[2] Abdeldjalil KHELASSI, Mohamed Amin Chick. Fuzzy knowledge-intensive case based classification for the detection of abnormal cardiac beats. Electronic Physician, 2012;4(2):565-571.

[3] KHELASSI, Abdeldjalil. Data mining application with case based reasoning classifier for breast cancer decision support. Proceedings of MICIT, Liverpool, UK, 2012.

[4] KLIR, George et YUAN, Bo. Fuzzy sets and fuzzy logic. New Jersey : Prentice Hall, 1995.

[5] ZADEH, Lotfi A. Fuzzy sets. Information and control, 1965, vol. 8, no 3, p. 338-353.

[6] HAN, Min, CAO, Zhanji, et LI, Yang. An improved case-based reasoning method based on fuzzy clustering and mutual information. In : Intelligent Control and Information Processing (ICICIP), 2014 Fifth International Conference on. IEEE, 2014. p. 293-300.

[7] LI, Hui, YU, Jun-Ling, YU, Le-An, et al. The clustering-based case-based reasoning for imbalanced business failure prediction: a hybrid approach through integrating unsupervised process with supervised process. International Journal of Systems Science, 2014, vol. 45, no 5, p. 1225-1241.

[8] KHELASSI, Abdeldjalil. Explanation-aware computing of the prognosis for breast cancer supported by IK-DCBRC: Technical innovation. Electronic Physician, 2014, vol. 6, no 4, p. 947.

[9] DALAL, Surjeet, JAGLAN, Dr Vivek, et SHARMA, Dr Kamal Kumar. Integrating Multi-case-base-reasoning with Distributed case-based reasoning. International Journal of Advanced Research in IT and Engineering, 2014, p. 2278-6244.

[10] LIU, Baoding. Uncertainty theory. Springer, 2014.

[11] KANAL, Laveen N. et LEMMER, John F. (ed.). Uncertainty in artificial intelligence. Elsevier, 2014.

[12] VLĂDUȚESCU, Ștefan. Uncertainty Communication Status. International Letters of Social and Humanistic Sciences, 2014, no 10, p. 100-106.

[13] HUGONNIER, Julien, MALAMUD, Semyon, et MORELLEC, Erwan. Capital supply uncertainty, cash holdings, and investment. Review of Financial Studies, 2014, p. hhu081.

[14] RICHTER, Michael M. et WEBER, Rosina. Case-Based Reasoning. Springer, Heidelberg, 2013.

[15] Abdeldjalil KHELASSI, '' Artificial Reasoning Systems: Theory and Medical Applications'' LAMBERT ACADEMIC publishing -LAP-, Saarbrücken, 2013.

[16] MARLING, Cindy, MONTANI, Stefania, BICHINDARITZ, Isabelle, et al. Synergistic case-based reasoning in medical domains. Expert systems with applications, 2014, vol. 41, no 2, p. 249-259.

[17] Abdeldjalil KHELASSI, '' Distributed Case-Based Reasoning Classifier for Cardiac Arrhythmias: Health sciences application'' LAMBERT ACADEMIC publishing -LAP-, Saarbrücken, 2012.

[18] WEISS, Gerhard (ed.). Multiagent systems: a modern approach to distributed artificial intelligence. MIT press, 1999.

[19] WOOLDRIDGE, Michael. An introduction to multiagent systems. John Wiley & Sons, 2009.

Extraction of Significant lexical Associations to Classify the Essential Perspectives for Text Summarization

N.Asadi[1], K.Badie[1], M.T.Mahmoudi[1] and N. Sahabi[2]
[1]Knowledge Management & E-Organization Group, Multimedia Group, IT Research Faculty, ICT Research Institute, Tehran, Iran
[2]Dept. of Algorithm & Computation, Faculty of Engineering & Science, University Of Tehran, Tehran, Iran

Abstract - *In this paper, a text summarization algorithm is proposed based on classifying the essential perspectives in texts. The algorithm is concerned with extracting significant lexical associations as the features for classification. Experimental results show that these features are enough to classify a number of pre-defined perspectives with a high frequency. In order to determine the related perspective, flow of consecutive sentences was shown to be most necessary.*

Keywords: Text summarization, perspectives, key phrase, lexical associations, sentence extraction.

1 Introduction

Due to the rapid growth of texts/ documents on the internet, accessing to the focal information within these documents has become significant. Within this scope the Information retrieval systems and search engines are at the user's service to find their required information. However, having retrieved the relevant documents, users should make sure of their usefulness and, very often, need to summarize the selected documents, which itself is usually a difficult and long-term procedure. Taking this point into account, automatic text summarization can play a promising role for such a purpose.

Text summarization is a process that takes a document as input and generates a shorter one, which is to represent the main content of the text, as output [1]. The summarized texts can be extractive or abstractive; the extractive summary is constructed by selecting important sentences of the document, while an abstract describes the content of the document in terms of sentences that do not necessarily appear identically in the initial document [2]. A document can be summarized from different points of view, depending on user's needs and knowledge.

The aim of this paper is to propose a new approach which may help us in generating a summary based on user's perspectives. The contribution of this paper is to propose a method to classify the sentences of a given paper in terms of some pre-defined perspectives. The following five perspectives have been selected: *Background, Related Works, Proposed Approach, Results and Future Works*. Regarding this, we make use of a machine learning algorithm according to which a number of features from each sentence, that are then applied to train a classification model, is extracted. Our classification achieved the accuracy of 88.7%, confirming the efficiency of the proposed features to classify the perspectives.

The paper is organized as follows: a literature review is presented in Section 2; Section 3 describes the proposed system and the evaluation and experimental results appear in Section 4. Section 5 is the concluding part of the paper.

2 Related Work

According to different categorizations, there exist different extractive and abstractive approaches to summarization from single to multi documents. Based upon these approaches, several characteristics can be considered for summarization, out of which frequency-based, characteristics, knowledge-based/ discourse-based characteristics, processing on surface, entities and discourse levels, can be enumerated. Besides the characteristics and levels of processing, a high attention should be paid to kinds of information through the spectrum of lexicon, structure information and matter of deep understanding, which are to be considered for summarization purposes [3,4,5].

As in this paper the focus of summarization is on lexicon/ keyword extraction, some of the existing approaches in this regard are discussed.

In some summarization approaches based on keyword extraction, the following phases are often seen: (i) converting the unstructured text into structured form and removing the stop words, parsing the text and assigning the POS (tag) for each word in the text and finally storing the result in a table, (ii) extracting the important key phrases in the text by some algorithms through ranking the candidate words, selecting the important sentences by using the extracted keywords/key phrases and measuring the similarity between the title and these sentences, (iii) extracting the sentences with the highest rank, and finally (iv) filtering and reducing some sentences in order to produce a qualitative summary [6].

Some other approaches also exist based on extracting keywords on the ground of lexical chains wherein the semantic similarity between terms are firstly calculated, based on which keywords are then extracted, and finally according to the lexical chain's intensity, entropy and position, the weighting importance of sentences are calculated [7]. In this regard, giving higher weights to words in the full-texts, as unigrams to be extracted as keywords, may improve the performance [8].

There are also some other research works which refer to producing extractive summaries of documents in the Kannada language. In the related algorithm, keywords are extracted from pre-categorized documents collected from online resources. To perform this process, firstly, features are obtained from documents, then scores obtained by GSS (Galavotti, Sebastiani, Simi) coefficients, IDF (Inverse Document Frequency) along with TF (Term Frequency) are combined to extract the keywords. This score of combination can later be used in ranking the sentences. At last, based upon the number of sentences given by the user, a summary would be generated [9].

Generating a summary by extracting sentence segments is also another method which is applied for summarization purposes. In this method, first, sentences are broken into segments by special cue markers. Then, a set of predefined features (e.g. location of the segment, average term frequencies of the words occurring in the segment, number of title words in the segment, and the like) are determined for each segment. Finally, a supervised learning algorithm is used to train the summarizer to extract important sentence segments, based on the feature vectors [10].

Baralis [11] proposed GRAPHSUM, a novel graph-based multi-document summarizer, which discovers frequent item sets, as correlations among multiple terms, and makes a graph-based model. The graph's nodes are combinations of two or more terms where its edges measure the strength of the associations between a pair of term sets. PageRank indexing algorithm is used to evaluate the relevance between graph's nodes, and the final summary will consist of the subset of sentences that best cover the indexed graph.

Beside the above mentioned methods, query-focused multi-document summarization based on keyword extraction is also mentionable. In this method, query related and topic related features for every word in the relevant document set are calculated. The importance degree of the words is then determined through combining these two features and the total score of candidate sentences based on their important words are respectively computed. At last, the candidate sentences with the highest score are selected as the summary sentences [12].

3 Proposed Approach

3.1 Basic Idea

Most of the existing text summarization systems are extractive, with no attention toward the document context and user's need [11,13,14]. Sentences are usually ranked by their scores so that a sentence with a higher rank can be of higher importance. However, a document can be summarized from different aspects. This may be confirmed by taking a look at summarizations made by a variety of people, each having made his work from his own point of view.

In this article, we propose a new approach which may help us, probably in the future to generate a perspective-oriented summary of a scientific paper. The main aim of this paper is to propose a method to classify the sentences of a given paper in terms of some pre-defined perspectives. Regarding this, we make use of a machine learning algorithm according to which a number of features from each sentence is extracted based on which classification model is trained.

Five perspectives, including *Background, Related Works, Proposed Approach, Results* and *Future Works* were selected due to their frequent usage in scientific papers which have been in practice written by a wide range of authors [15]. Table 1 illustrates brief description of these perspectives.

3.2 Essential phases in the proposed approach

3.2.1 Preliminaries

Our purpose is to take a scientific paper as the input and classify its sentences into some predefined perspectives. In the first phase, the papers are preprocessed to extract those sentences which correspond to each perspective. In the next phase, a weight is assigned to all sentence features, and the training data will then be generated. Finally, in the last phase, the feature vectors are classified in terms of the predefined perspectives. A pseudo-code of the proposed approach is shown in Figure 1.

Table 1: Description of perspectives

Perspective	Related Subtitles
Background	Introduction, Background
Related Works	Related work, Literature review, Previous work
Proposed Approach	Proposed Approach, Proposed Method, Approach
Results	Evaluation, Experimental results, Experiments
Future Works	Future Work, Conclusion, Discussion

3.2.2 Preprocessing

In this step, the papers are split into separate files, taking into account the subtitles of the sections, and their keywords will also be stored in separate files. Having extracted the sentences from each section, short sentences will be removed. Here, we use POS tagging and lemmatization to recognize verbs and nouns in the sentences.

3.2.3 Feature weighting

Many statistical and linguistic features are in reality considered for classification of sentences in these perspectives, six of which were finally selected amongst. We are going to explain these perspectives more specifically.
An empirical study on many papers convinced us that the typical words used in any particular section may, to some extent, determine the title of that section. This is because the type of verbs and phrases appearing in a paper vary, very often, from one section to another. For instance, general verbs are usually used in *Background* while in *Proposed Approach*, special verbs are more frequent. Also in *Related works* many special phrases (e.g. the name of methods and tools) may show up. In contrast, special terms are used seldom in *Results*. Taking these into account both general and special verbs as well as phrases were decided to constitute the feature vector. The number of keywords and citations, as auxiliary features, were also decided to be added to the feature vector as suplementory information for separating the pre-defined perspectives more accurately. Numerous citations and keywords are also believed to appear well in the *Related Works* but not so freqently in results..
More details of the selected features are illustrated in Table 2.

Table 2: Description of Features

Features	Description	Example
Special Phrases	The number of phrases that are used in a specific domain, for example the name of methods and tools	Lemur toolkit
General Phrases	The number of phrases that are used in most of the papers	Information retrieval system
Special Verbs	The number of verbs and actions that are used in specific domains	Mine, Parse
General Verbs	The number of verbs and actions that are used in most of the papers and arrange the text.	Explain, describe
Citation	The number of references in each sentence	(Radev, 2008)…
Keywords	The number of keywords in each sentence	

To extract the features explained in Table 2, some points are considered as follows:

(i) Phrases Extraction: We extracted important phrases of training data and classified them into two group of *special phrases* and *general phrases*. These phrases are separately stored in two hash tables.
The following two patterns have been considered for extracting phrases from sentences.

$$Noun\,(Noun \mid Adjective)* \qquad (1)$$

$$\text{Adjective (Noun|Adjective)}^{+} \qquad (2)$$

The phrases may be fully and also partially searched within the hash tables. For example, besides "Boolean Information retrieval", "Boolean Information", "Information retrieval", "Boolean", "Information" and "retrieval" are also searched

(ii) Verbs Extraction: The verbs being used in the training data are labeled as *special verbs* or *general verbs*, which are separately stored in two hash tables. After POS tagging, the verbs of the sentences are searched within the hash tables. The frequency of each class will then be used to evaluate the feature of special or general verbs.

(iii) Citation Extraction: We use a pattern matching technique to find the number of all citations in each sentence. The three reference patterns "[reference number]", "(author, year)" and "author (year)" were considered in this regard.

For every sentence belonging to each perspective, the listed features are extracted. A window with size of n, slides over the sentences and stores the sentences appearing in the window as a piece of the text. Here, the number of windows generated from a section consisting of k sentences is (k-n+1) and the weight of a particular feature of the window is the sum of the weights in the corresponding in the window "(3)".

$$\text{Featuer}_{iw} = \sum_{s=1}^{n} \text{Feature}_{is} \qquad (3)$$

where Feature_{iw} is the i-th feature in window w and Featuer_{is} is the i-th feature in sentence s.

Note that for an instance of a feature, its class label corresponds to the perspective within which the window is located.

3.2.4 Classification

Having produced the feature vectors, we use the sequential minimal optimization (SMO) algorithm with polynomial kernel to learn the training data. This is because the SMO is capable of handling multi-class classification and that the polynomial kernel could help us to separate pre-defined perspectives in a suitable way. Also, to evaluate the

accuracy of the classification, 10-fold cross validation is applied.

SMO algorithm, introduced by Platt [16] for training Support Vector Machines, is the fastest algorithm applicable on linear SVMs and sparse data sets. For polynomial SVMs, it is as fast as PCG chunking. SMO may handle very large training sets since it requires linear memory. We used SMO implementation in WEKA, during which pair-wise coupling; a strategy for multi-class classification including estimating class probabilities for each pair of classes while coupling the estimates together [17], was applied. SMO replaces all missing values and transforms nominal attributes into binary ones.

```
1: procedure PERSPECTIVECLASSIFICATION(ScientificPapers , fold , n)
2:     Split ScientificPapers into seperate files regarding the subtitle of the
   sections (perspectives)
3:     for i = 1 to perspectives do
4:         while has more documents in perspective i do
5:             for all sentences s ∈ document j do
6:                 POSTAGGING(s)
7:                 CREATEFEAUREVECTOR(s)
8:             end for
9:             Slide a window w with size of n over the sentences ∈ document j
10:            for all sentences s ∈ window w do
11:                FeatueVector_w+ = FeatureVector_s
12:            end for
13:        end while
14:    end for
15:    ClassificationModel ← SMO ALGORITHM(FeatureVectors)
16:    CROSSVALIDATION(ClassificationModel , fold)
17: end procedure
```

Figure 1 : Pseudo code of proposed approach

4 Experimental Results

4.1 Dataset:

We made use of 20 papers from IEEE, concerning information retrieval and data mining as our data set, the sections of each of which were split into a separate file. Total number of sentences in the dataset is 4516. Table 3 gives the number of sentences for each perspective.

Table 3: Number of sentences in each perspective

Perspective	Number of sentences
Background	657
Related Works	654
Proposed Approach	1742
Results	1219
Future Works	244

4.2 Determining the size of the window

To determine an efficient value n for the number of sentences located on a window, we run our proposed algorithm with different values of *n*. The results of the classification accuracy are shown in Figure 2.

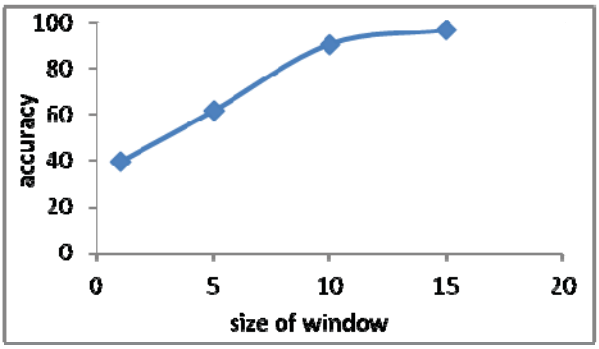


Figure 1: Classification accuracy for different window sizes

Note that the higher the value of *n*, the more classification accuracy is obtained. When *n* is considered as the number of sentences in a particular section, the accuracy will be nearly 100%. This is clear because the whole part of a section reflects a perspective in a more desirable way compared to the case of few sentences. However we set *n=10*; this is somehow an optimal value, since lower values may lead in a decrease in accuracy and larger values lead in significantly long windows which will not be efficient to consider and, moreover, substituting 10 by much greater numbers will no longer make any significant improve in accuracy.

Instead of a single sentence, we need to consider a flow of some consecutive sentences in a text to determine the related perspective.

4.3 Evaluating the Performance of Sentence Classification

Using the five predefined perspectives, we labeled our dataset sentences and applied then SMO to learn them. 10-fold cross validation was also used to evaluate the performance of sentence classification. Our classification achieved the accuracy of 80.4% for *n=10*. The details are shown in Table 4 below.

Table 4: Classification results

Class	Precision	Recall	F-measure
Background	0.85	0.711	0.784
Related Works	0.79	0.89	0.844
Proposed Approach	0.797	0.772	0.788
Results	0.848	0.703	0.76
Future Works	0.96	0.409	0.574

results obtained indicate that the extracted features are capable enough to classify our perspectives in a suitable way. *Background, Proposed Approach, Results* and, particularly, *Related Work* as perspectives have been classified with high precision and recall. The recall of *Future Works* class is low because the number of sentences which have been labeled as *Future Works* is low, and besides that the classifier at hand has not been trained sufficiently.
In another experiment, phrases and verbs are divided into three groups and are labeled as *special, general* and *speciogeneral*. The latter includes those phrases which are used to represent an action or a well-known algorithm, i.e., "classification models", "Information retrieval", etc. More specifically, the afore-mentioned *general phrases* are classified as *general* and *speciogeneral phrases*. The labeling of the verbs here differs a bit in the following sense: those verbs which might be used both as *general* and *special*, have been considered to be *speciogeneral* verbs, depending on the topic of the paper.

In order to evaluate the effect of new features, the proposed algorithm has been evaluated on some combinations of features. In this regard, the following four cases have been considered in the experiments:
1) Main features
2) Main features + *speciogeneral verbs*
3) Main features + *speciogeneral phrases*
4) Main features + *speciogeneral verbs* + *speciogeneral phrases.*
A summary of the results has been shown in Table 5.

Table 5: classification results with different combinations of features

Features	Accuracy
Main features	80.4%
Main features+ sog verbs[1]	84.9%
Main features + sog phrases[2]	85.6%
Main features + sog verbs + sog phrases	**88.7%**

As indicated in Table 5, the positive impact of both *speciogeneral verbs* and *speciogeneral phrases* is clear.

Classification details show that, using *speciogeneral verbs*, *Future Works* and *Results* classes are better separated and *Results* are appropriately distinguished from *Proposed Approach*. Using *speciogeneral phrases*, the accuracy of classification between *Related Works, Proposed Approach* and *Results* has been increased, and an improvement in the performance of the classification of *Background* and *Future Works* is also tangible.

[1] Speciogeneral verbs
[2] Speciogeneral phrases

5 Conclusions

In this paper, we proposed an algorithm to classify the sentences of a scientific paper into some predefined perspectives, namely *Background, Related Works, Proposed Approach, Results* and *Future Works*. Our algorithm involves three phases. The first phase is for preprocessing the content f papers to extract the sentences corresponding to each perspective. In the second phase, a weight is assigned to every feature of the sentences and, finally, in the last phase, the feature vectors are classified into predefined perspectives by SMO algorithm. The number of citations, keywords, phrases and verbs in each sentence have also been considered as supplementary features. Experimental results show that the selected features are capable enough of classifying the desired perspectives with a high accuracy. As future works, other classification methods like Random Forest, k*, etc. can be exploited so that their results may be compared to those of the SMO. Results also indicate that, instead of a single sentence, we need to consider a flow of some successive sentences in order to determine the corresponding perspective.

Labeling of verbs and phrases into the three groups of *special*, *general* and *speciogeneral*, which is performed manually depending partly on one's particular ideas, is one of the limitations of the current approach which may lead to some problems with regard to summarization. Automating such a process can therefore be viewed a promising research work in future.

6 References

[1] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications*. Amesterdam/Philadelphia: John Benjamins Publishing Company, 2002.

[2] L. Elena, M. T. Roma-Ferri, and M. Palomar, "COMPENDIUM: A text summarization system for generating abstract of research papers," *Data & Knowledge Engineering*, pp. 164-175, 2013.

[3] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," *Journal of Computer Science*, vol. 8, no. 1, pp. 133-140, 2012.

[4] A. Nenkova and K. McKeown, "A Survey of Text Summarization," in *Mining Text Data*. Springer, 2012, pp. 43-76.

[5] L. Suanmali and N. Salim, "Literature Reviews for Multi-Document Summarization," 2008.

[6] R. Al-Hashemi, "Text Summarization Extraction System (TSES)," *International Arab Journal of e-Technology*, vol. 1, pp. 164-168, 2010.

[7] J. Xiao-Yu, "Chinese Automatic Text Summarization Based on Keyword Extraction," in *First International Workshop on Database Technology and Applications*,

2009, pp. 225-228.

[8] A. Hulth and B. M. Beata, "A study on automatically extracted keywords in text categorization," in *Proceeding of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.*, 2006.

[9] R. Jayashree, K. M. Srikanta, and K. Sunny, "Keyword Extraction based Summarization of Categorized Kannada Text Documents," *International Journal on Soft Computing(IJSC)*, vol. 2, no. 4, pp. 152-159, 2011.

[10] W. T. Chuang and Y. Jihoon, "Extracting sentence segments for text summarization: a machine learning approach," in *Proceeding of the 23rd annual international ACM SIGIR conference n Research and development in information retrieval.*, 2000.

[11] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, "GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization," *Information Sciences*, vol. 249, pp. 96-109, 2013.

[12] L. Ma and e. al, "Query-focused multi-document sumarization using keyword extraction," in *INternational Conference on Computer Science and Software Engineering*, vol. 1, 2008.

[13] R. Ferreira, L. d. S. Cabral, R. D. Lins, G. P. e. Silva, and F. Feritas, "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, vol. 40, pp. 5755-5764, 2013.

[14] K. Nandhini and S. R. Balasundaram, "Improving readability through extractive summarization for learners with reading difficulties," *Egyptain Information Journal*, pp. 195-204, 2013.

[15] K. Badie, M. Kharrat, M. T. Mahmoudi, and S. Miran, "Ontology-Driven creation of contents: Making efficient interaction between organizational users and their surrounding tasks," in *User Interface Desgin for Virtual Environment: Challenges and Advances*, B. Khan, Ed. Heshey, PA: Information Science Reference, 2012, pp. 156-170.

[16] J. Platt, "Fast training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods-Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[17] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637-649, 2001.

post-*Mayo/Biosig/Alice* – The Precise Meanings of Their New SPL Terms

Sigram Schindler[1,2], Bernd Wegner[1,2], Juergen Schulze[2] and Doerte Schoenberg[2]
[1]Technische Universität, Berlin, Germany
[2]TELES Patent Rights International GmbH, Berlin, Germany

Abstract - *This paper precisely defines the new and fundamental notions that the Supreme Court unanimously introduced into SPL precedents for ET CIs by its KSR/ Bilski/Mayo/Myriad/Biosig/Alice line of decisions, i.e. of the terms "scope"/"definiteness"/"preemptivity"/"natural phenomenon"/"abstract idea"/"categories of patent-non-eligibility". Hitherto, none of these notions was precisely understood, making SPL precedents on ET CIs error prone – see recent CAFC decisions, broadly criticized by patent experts at the USPTO event on patent quality, not aware of the semiotic necessities of adjusting SPL precedents to the needs of ET CIs.*

Keywords: "scope", "definiteness", "preemptivity", "natural phenomenon", "abstract idea", "categories of patent-non-eligibility"

1 THE post-*Mayo/Biosig/Alice* REFINED NOTIONS OF SPL[1.a)] PRECEDENTS FOR ET CIs

This paper precisely defines the new and fundamental notions[1.b)2.a)] that the Supreme Court unanimously introduced into SPL precedents for ET CIs by its *KSR/ Bilski/Mayo/Myriad/Biosig/Alice* line of decisions, i.e. of the terms "scope"/ "definiteness"/"preemptivity"/"natural phenomenon"/"abstract idea"/"categories of patent-noneligibility". Hitherto, none of these notions was precisely understood, making SPL precedents on ET CIs error prone – see recent CAFC decisions [163], broadly criticized by patent experts at the USPTO event on patent quality [193], not aware of the semiotic necessities of adjusting SPL precedents to the needs of ET CIs.

FIGs 1&2 are key for grasping the many strong interdependencies SPL imposes on these notions: Any one of them is defined only if all notions it "uses" are[2.b)]. A statement on one of these notions and ignoring such an interdependency is flawed. The reasonability of the crucial notion "abstract idea" will be made evident. In total, for the first time all these notions will be precisely defined and any question about them clearly answered – enabled by the Supreme Court's above line of decisions.

FIG 1: The Subtests Used in the Classical and in the Refined Claim Construction

FIG 2: The Semi-Automatic FSTP^{FFOLLIN}-Test of a CI's TT0 – required for ET CIs.

As to [183], FIG 1 is left unchanged and FIG 2 only slightly refined[2.a) 3.a)].

---

[1] **a.** SPL = Substantive Patent Law = 35 USC §§ 101/102/103/112; ET/CT = Emerging/Classic Technology; CI = Claimed Invention; NPS = National Patent System. This paper continues [183], not considering these notions in any other context, and in particular not "per se"/"as such", as this were totally irrational.

.b A "term" is an "identifier/name". A pair <"term", its "meaning"> is called the term's "notion".

[2] **.a** All notions are defined, for a TT0, by **1.)** assuming it had passed the whole FSTP-Test, **2.)** deriving these TT0 notions' precise meanings (in mathematical KR) from the so achieved TT0 presentation by its BAD/BED-inCs [183]. Without **1.)**, these notions are only "indicative"/"intentional", i.e. not defined/-able.

In Physics it is usual to perform such "retrospective"/"fiction based" definitions: There – for finding out what a system's properties in its equilibrium state are – one always **1.)** assumes it had already reached the equilibrium state, **2.)** determines, in this state, relations between its constituents. Taking a TT0 as being the analogon to a physical system, its analogon to the latter's **α)** equilibrium state is that it has passed the whole FSTP-Test, i.e. satisfies SPL, and **β)** relations between its constituents are TT0's relations between its inCs and these notions. An even closer such analogon – as also "sub-Physics", just as SPL notions, "SPL loaded" instead – exists in Mathematics' foundation, i.e. in Measure Theory [191].

All above notions are defined by assuming the BAD/BED-inC instantiations are known. In a specific TT0's SPL test their values are needed. This usually will require reiterating the determinations of all these instantiations until they all are consistent. Thereby not only the currently performed claim construction is readjusted but even the claim interpretation preceding it [183]. Also the refinements/disaggregations may be reiterated – here assuming the product of their domains may not be disaggregated into a an equivalent but non-isomorphic way. The next Sections show: Without reiterations S and S^R for a TT0 – and relating them to their definitions[6.a)] – this consistency is hardly achievable.

.b There is not only a "use-hierarchy" – defined by David Parnas [122] – between the FSTP-testi's, as it here is reasonably assumed for efficient execution of the FSTP-Test. The reason being: SPL also imposes the inverse relation on these testi's: Thus, for the above notions defined by the test.i's, SPL (in the Supreme Court's interpretation by its above line of decisions) clearly implies[2.a)] that, for a TT0, "any of these notions is defined only if all such notions are defined", too – explaining why one cannot prove FSTP-testi's holding independently of ∀ FSTP-testi^{o}'s holding, $1 \leq i^{o}, i \leq 10$, repeatedly reminded of below.

2 PRECISELY DEFINING THE NOTIONS OF "SCOPE" & "DEFINITENESS"

While the ambitions of the FSTP-Project in total reach out very far – at developing an extremely powerful IES [161], for which an in SPL precedents hitherto unknown/non-practiced notional preciseness is indispensable – here this preciseness is introduced and its implied necessary subtlety of reasoning is explained. This is done by means of the FSTP-Test, FIG 2, also being the backbone of the IES.

The following elaborations on SPL encounter legal questions – below identified by "**(!!)**" and not yet settled by SPL precedents, as not so precise about these new notions – which must be answered for ET CIs here for not getting blocked by them. These answers should be in line with the Supreme Court's above decisions[3.a)].

The FSTP-test1 prompts the user to input, for CI/TT0[2.c)] from doc0, ■) its "CI- elements, X0n" , 1≤n≤N, ■) their by mathematical predicates modeled compound inventive concepts BAD-crC0n, ■) as many elementary inventive concepts BED-crC0nk as it is able to identify[2.a)] $1\leq kn\leq K^{n}$, $K::=\sum^{1\leq n\leq N}K^{n}$, which define CI's sole[2.c)3.a)] set $S=\{s^{k}|1\leq k\leq K\}$ – therein identified all BAD-crC0n***** & BED-crC0k***** subject to a patent-noneligibility exemption – and ■) all justifications prompted for on lines 1)(b)-4) in FIG 2. After clarifying the above quoted notions, the RS and FSTP-test9/10 –though relevant[2.b)] – no longer need additional clarification, here [182].

D.1: $S^{R}::=\{\forall s^{Rv}\}::=\{\forall<s^{Rv1}\epsilon TS(s^{1}),..,s^{RvK}\epsilon TS(s^{K})>\}$ is called "TT0-REALIZATION SET" iff $\forall s^{Rv}$ the "s^{Rv}-embodiment, $TT0^{sRv}$" is disclosed by TT0's specification.[2.a)3.c/d)]

LEMMA: For TT0 – by the independency of its BED-inC0kn by FSTP-test3 – holds $S^{R}\cong\prod^{\forall sRv}s^{Rv}\cong^{3.c)}\prod^{1\leq k\leq K}TS(s^{k})\cong\prod^{1\leq n\leq N\wedge 1\leq kn\leq Kn}TS(\text{BED-inC0kn})$.

D.2 "SCOPE(TT0)": S^{R} is called "scope(TT0)", resp. "scope(CI)" iff ∃ only 1 TT0[3.a)].

This is the first time that this fundamental notion of SPL – the scope of a TT0 – is precisely defined. Hitherto, namely nobody had developed the notion of a TT0's "realization set"[D.1], being its "protected embodiments set" – decisive for TT0's alleged infringement[D.5]. If for an s^{Rv} the TT0 specification discloses of its $TT0^{sRv}$ for the posc no enablement, then it is impossible to justify in FSTP-test5 this S satisfies SPL – as this were the just indicated legal error – and measure[3.c)] is to be taken.

In "classical" claim interpretation/construction [183] thus hitherto a legal deficiency was principally absolutely unavoidable: To assess TT0's enabling disclosures by its specification not of all but just of a few TT0 embodiments – assuming the posc then would understand them all, without being capable of saying, what for TT0 exactly this "all" would comprise!!! This question arises in any infringement dispute and could hitherto never be answered precisely – quite principally!!!

This also caused the "overclaiming" problem of a CI – meaning that its claim is disclosed "overbroad", thus being strong in talking a similar invention into infringing it, but untenable in its defense against its nullification as deficient as to its complete enablement. This now is easily avoidable by obeying: $\text{scope(CI)} = \{\forall\ TT0^{sRv}\}$[3.c)].

Legally, S^{R}(TT0) is the scope of TT0's protection by patent law if and only if ("iff") TT0 has passed the complete FSTP-Test[2.a)] – otherwise scope(TT0) is not defined at all, and there is no protection for TT0 by patent law. [2.a)4.a)].

D.3 "TT0' = TT0"[4.b)]: A TT0' is called to be "equal, '=' " to TT0 iff $S'^{R}=S^{R}$.

---

[3] **.a** But these also are the simplest answers – hence, potentially too rigorous in practical cases. As soon as ∃ precedential decisions, deviating from this rigor, the following definitions may need marginal modifications – fortunately changing nothing of principal significance, as recognizable today already[3.a)v.].

Five notes concerning abbreviations (used below as already in [183]) and the precision ahead:

i. The index "FFOLLIN" is omitted here, as in the FSTP-Test after its preamble. But it should be kept in mind: The below insights apply to many other Intellectual Property Laws/Regulations, too.
ii. Throughout this paper is assumed, a CI has just 1 interpretation/"Generative Set, S"/TT0 [58]. This restriction may be dropped by applying the FSTP-Test to all finitely many TT0s alias S's of CI.
iii. Some sloppy wordings of [183] are fixed. Note also: Talking about TT0' assumes TT0 ∃ already.
iv. For preciseness, definitions – abbr. by D.i, i=1,2,… – use (basic) Mathematics.
v. Independently of risks with future SPL precedents, scientifically the here defined answers are well defined and hence will prevail – potentially identifying the difference to SPL precedents.

.b FOL enables: $\forall n\epsilon[1,N]\wedge\forall k\neq k'\epsilon[1,K^{n}]$ holds BED-inC0kn≠BED-inC0k'n. Also the simplification is assumed that, if ∃ several BED-inC instantiations in an s^{Rv}, they all have the same value. Due to all sets' finity, all such[ii.] simplifications are removable by expanding the FSTP-Test to remain exhaustive.

.c By appropriate inC limitations, the set equality "≅" may be preserved also if some s^{Rv}'s are not disclosed, i.e. are $\notin S^{R}$ – whereby this reduced S'^{R} evidently represents a TT0'≠ TT0[6.a)], see D.**2**-D.**5**. I.e.: In spite of the independency of TT0's BED-inCs[2.a)], the definitions of their TSes may impact on each other.

.d Analogous terms/notions $\mathbf{S'^{R}}$, s'^{Rv}, $\prod^{1\leq k\leq K}TS(s'^{k})$ are used also for a TT0', which need not pass the FSTP-Test (e.g. because there is no TT0-alike TT0' specification or inC definition), i.e. for which only little of TT0 holds – whereby in any D.i its TT0' notions are used like TT0 notions.

[4] **.a** TT0, TT0'ϵFFOL => $|S^{R}|,|S'^{R}|$ are finite, i.e. for both there is no FSTP-Test termination problem.

.b It were false to conclude, in D.3, TT0' = TT0, just because $\prod^{1\leq k\leq K}TS(s'^{k})=\prod^{1\leq k\leq K}TS(s^{k})$, i.e. without verifying that TT0' passes also the whole FSTP-Test, as ∃TT0 ∧ ∃TT0' ∉ scope(TT0) **:** TT0' meets this "product = requirement". E.g., $\{s'^{k}\}$ need not be independent or well-defined over posc, i.e. not meet FSTP-test3/-test4.

It were false to conclude, in D.4, TT0' is ϵ scope(TT0) just because $\prod^{1\leq k\leq K}TS(s'^{k})\leq\prod^{1\leq k\leq K}TS(s^{k})$, i.e. without verifying that TT0' passes also the whole FSTP-Test, as ∃TT0 ∧ ∃TT0' ∉ scope(TT0) **:** TT0' meets this "product ≤ requirement". E.g., for some TT0 simply define TT0' by removing from S^{R} an s^{Rv}, as discussed above[3.c/.d)].

.c In D.**5** suffices already: $\exists s'^{Rv}\epsilon\ S'^{R}\cap S^{R}\wedge$TT0' not passes the FSTP-Test[3.d)] => TT0' violates scope(TT0).

.d This "TT0 is definite" definition D.**6** is equivalent to *Biosig's*, but needs no unknown CIs (like *Biosig* does). I.e.: A TT0's definiteness test is part of TT0's FSTP-Test, i.e. comes for free. The conclusion is, the Biosig test is equivalent to the FSTP-Test, but just declarative, i.e. not operational, as the FSTP-Test. A (nontrivial) equivalence proof is: By D.**4** holds: TT0'ϵ scope(TT0) ⇔ TT0' passes the FSTP-Test $\wedge S'^{R}\subseteq S^{R}$.

D.4 "TT0' ϵ SCOPE(TT0)"[4.b)]: A TT0' is called to "belong to scope(TT0)" iff TT0' passes the FSTP-Test ∧ $S'^{R} \subseteq S^{R.}$
D.5 "TT0' VIOLATES TT0" A TT0'∉SCOPE(TT0) is called to "violate" TT0 iff $S'R \cap SR \neq \Phi$[4.c)].

D.6 "TT0 IS DEFINITE"[4.d)]: A TT0 is called "definite" iff it passes the FSTP-Test.

Finally: from D.4 ∧ D.6 trivially follows: TT0' ϵ scope(TT0) => TT0' is definite. [5.a)]

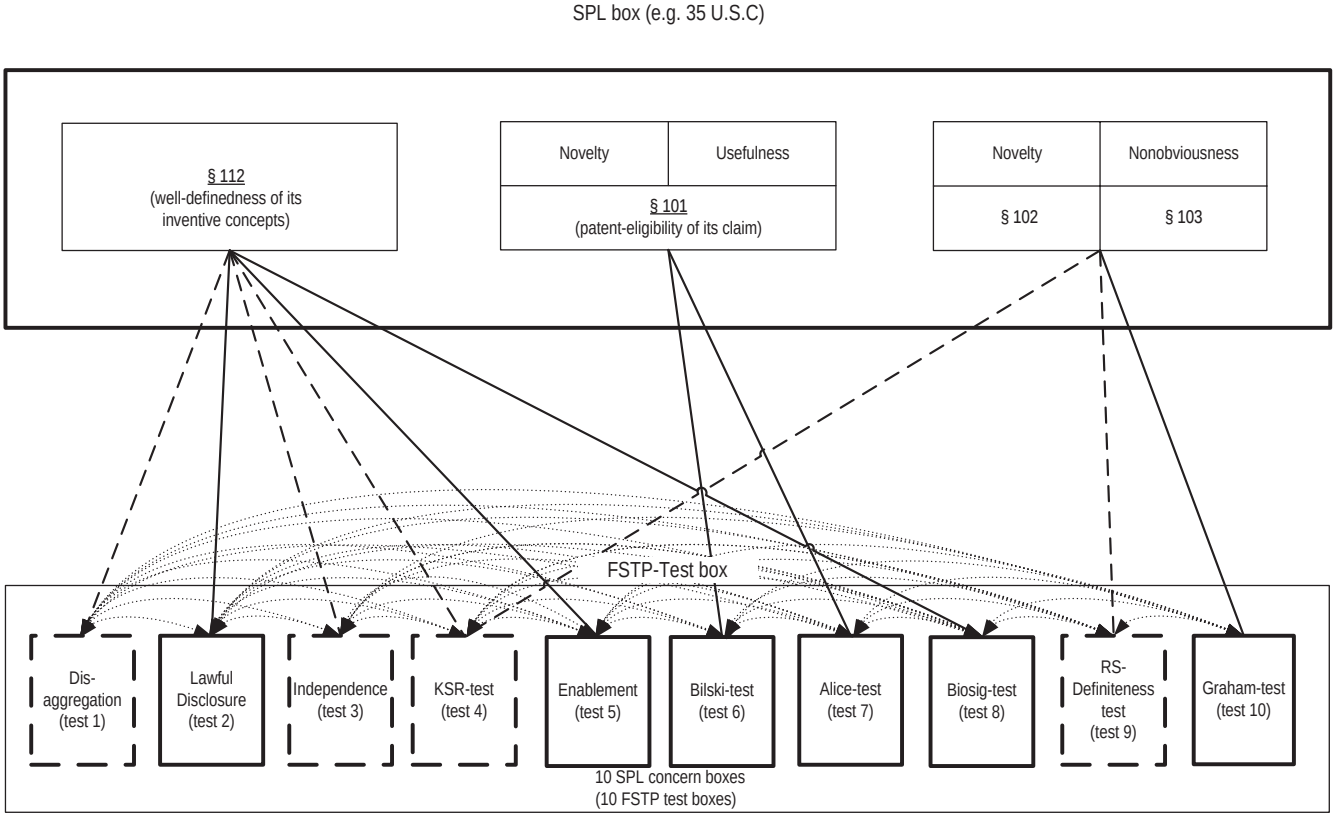


FIG 1

Bold lines show the classical claim construction's test.i's, dashed ones what Mayo/Biosig/Alice additionally require (refined claim construction). "←" show a "use hierarchy" among test.i's. "→" expand it to testi's total dependency.

==========

The $FSTP^{FFOLLIN}$-Test is a computer implemented method – defining also a system – for testing

- under a given Finite First Order Logic Legal Invention Norm, FFOLLIN, a given Claimed Invention, $CI^{FFOLLIN}$, which has a given interpretation $TT0^{FFOLLIN}$, represented by its Generative Set of $TT0^{FFOLLIN}$, $S^{FFOLLIN}$,
- $TT0^{FFOLLIN}$ – defined by $S^{BADFFOLLIN}$ ::= {BAD-crC0n^{FFOLLIN},1≤n≤N} ∧
  ∧ $S^{FFOLLIN}$ ::= {BED-crC0kn^{FFOLLIN} | 1≤n≤N : BAD-crC0n^{FFOLLIN}=∧$^{1≤kn≤Kn}$BED-crC0kn^{FFOLLIN}},

whether this FFOLLIN is satisfied by $TT0^{FFOLLIN}$ alias $S^{FFOLLIN}$,

- whereby FFOLLIN is defined to comprise a conjunction of 10 given $FSTP^{FFOLLIN}$-test.o of $TT0^{FFOLLIN}$ alias $S^{FFOLLIN}$, i.e. ∧$^{1≤o≤10}$$FSTP^{FFOLLIN}$-test.o – for brevity in the sequel the index "FFOLLIN" being omitted, any FSTP-test.o abbr. by just "o)", 1≤o≤10, and for 6≤o≤10 the stereotypic "over model and posc" omitted –

whereby the claimed invention for any TT0 prompts the CI's user to input to it

- the given information ■) ∀TT0-elements X0n of TT0, 1≤n≤N, ∧ ∀ binary abstract and elementary disclosed creative concepts of all X0n, BAD-crC0n resp. BED-crC0n ■) for |RS|>0 also ∀TTi-(dummy-)elements Xin peer to X0n, 1≤i≤I=|RS| ∧1≤n≤N, ∧ ∀ binary abstract and elementary disclosed (dummy-)creative concepts, crCin, of all (dummy-)elements Xin, called BAD-crCin resp. BED-crCin, as well as ■) ∀ below justifications, by stepwise prompting,

i.e., for testing the S input to it as follows:

1) (a) S^{BAD}::={BAD-crC0n |∀$^{1≤n≤N}$}, S::={BED-crC0kn|1≤n≤N:BAD-crC0n=∧$^{1≤kn≤Kn}$BED-crC0kn};
   (b) justof$^{∀1≤n≤N}$: BAD-crC0n is definite,
   (c) justof$^{∀1≤n≤N∧∀1≤kn≤Kn}$: BED-crC0kn is definite ∧ ∀ patent-noneligible BED-crC0kn* are identified;
   (d) justof$^{∀ϵSBAD∪S}$: BAD-crC0n = ∧$^{1≤kn≤Kn}$BED-crC0kn;
2) justof$^{∀ϵSBAD∪S}$: sϵS ∧ BAD-crC0nϵS^{BAD} are lawfully disclosed;
3) justof$^{∀ϵSBAD∪S}$: Independence-test passed S is well-defined&independent over model;

4)	justof$^{\forall \epsilon SBADUS}$:	*KSR*-test passed	S is well-defined over posc;
5)	justof$^{\forall \epsilon SBADUS}$:	TT0's implementation by S is enablingly/lawfully disclosed;	
6)	justof$^{\forall \epsilon SBADUS}$:	*Bilski*-test passed	TT0 is non-preemptive;
7)	justof$^{\forall \epsilon SBADUS}$:	*Alice*-test passed	TT0 is patent-eligible;
8)	justof$^{\forall \epsilon SBADUS}$:	*Biosig*-test passed	TT0 is definite;
9)	justof$^{\forall \epsilon SBADUS}$:	RS-Definiteness-test passed	RS is well-defined over TT0;
10)	justof$^{\forall \epsilon SBADUS}$:	*Graham*-test passed	TT0 is patentable.

FIG 2:
The FSTP^{FFOLLIN}-Test, the passing of which is necessary and sufficient for a CI's TT0 to satisfy SPL

3 PRECISELY DEFINING THE NOTIONS OF "PREEMPTIVITY" & "NATURAL PHENOMENON" & "ABSTRACT IDEA" & "PATENT-ELIGIBILITY"

D.7: Induced by *Mayo* let, for a TT0's CI-element, the term "improvement-prone, ip" denote a new "property category" for its inC(s), modeled as its(their) "ip-inC(s)". Compared to such an inC, its new ip-inC property is: It is already 'visible' that it will "improve" in its domain and/or its TS, no matter whether predictably in time or not.

Of any CI and/or its inC(s), this definition of an ip-inC enables precisely modeling its/their natural phenomenon or abstract idea "(property) category"[6.b)] – so the *Alice* term[5.b)]. In principle, both categories make its CI preemptive – introduced in *Bilski/Mayo/Alice* and modeled below – i.e. they may be seen as categories of such "ip-CIs" exemption from patent-eligibility[5.b)]. This definition of ip-inCs yet enables them to themselves avoiding preemptivity and hence patent-noneligibility of the CIs embodying them (i.e. in their Generative Sets) – as explained below.

ip-CIs' specifications need not disclose enablements for their future potential improvements their ip-inCs model. I.e., for an ip-CI having its FSTP-Test would not perform FSTP-test5 for the improvements of its ip-inCs over their original inCs.

The ip-inCs' capability to precisely identify, right from the outset of applying for a patent for a CI its potential future improvements, evidently shall provide a clean framework for implicitly conveying, with this actual application, a preliminary patent application for the improved respective CI – as outlined below in some detail. The similar but less precisely defined effect has hitherto been achieved by simply mentioning potential future CI improvements in its specification. If this mathematical modelation framework is socially felt as being too narrowing, it may be pragmatically relaxed – then tolerating some preemptivity by SPL precedents using it. **(!!)**

Two currently broadly discussed ip-inCs – though not so termed:

1.) In the Supreme Court's *Myriad* decision the "BRCA" TT0 has a single CI-element (representing a specific chromosome of the human genome) and its single inC – an aspect of this chromosome modeled by domain(inC)– models, by its TS(inC), "a huge range of 'certain nucleotide sequences'". R&D will visibly but unpredictably in time improve, e.g. TS(inC). Hence, this inC is a natural phenomenon ip-inC.

2.) In *Alice*, its "Closing" CI-element of its "transaction settling" TT0 today has only a single inC ("automatically in the evening"), but its TS will utmost likely but unpredictably improve over time in response e.g. to customer demands (for an additional "close on request instantly", …). Hence, this inC is an abstract idea ip-inC.

D.8: For an $s \epsilon S$ and an s^{o} let be defined ■) the relation "$\mathbf{s^{o}>s}$" iff domain(s) = domain(s^{o}) ∧ TS(s^{o})\TS(s)≠Φ, and ■) as meaning of "s=ip" to be that s is an "ip-inC".

D.9 "PREEMPTIVITY" by *Bilski*: TT0 is called "preemptive" iff $\exists TT0' \neq TT0$ passing the FSTP-Test**:** $scope(TT0') \cap scope(TT0) \neq \Phi \wedge \exists k \epsilon [1,K]$**:** $(s'^{k} > s^{k}) \vee (s^{k}=ip)$[5.b)].

---

[5] **.a** D.1 introduced a notion reflecting the "scientific standard procedures" in analyzing a problem, it is the SPL-specific analog notion to the well-known notion of "state-diagram", common to all exact sciences.

D.2- D.6 introduced notions mirroring alias modeling the basic scientification impacts of the above quoted Supreme Court decisions on SPL precedents, indispensable for enabling it to consistently dealing with ET CIs. These notions are basic in that they deal with terms having semantics, which – also that of "inventive concepts, inCs" – ought to have been known in the pre-*Mayo/Biosig/Alice* era, already.

The fact that they have not been clarified earlier by SPL precedents tells that it hitherto has not been subject to the scrutiny of scientific analysis – although the Supreme Court by its above quoted decisions repeatedly implicitly asked for it.

D.7- D.14 precisely define further notions indispensable for enabling SPL precedents to consistently deciding on ET CIs in the light of the terms/notions of this Sections' headline. Their semiotics also models scientification impacts of the above quoted Supreme Court decisions on SPL precedents [171].

This paper considers the semiotics alias "SPL precedents new meaning making" for the terms/notions "ip-inCs" and "tw-inCs" introduced by D.7 (enabling clarifying the notion of "preemptivity" and "abstract idea") resp. D.11[5.b)] (on this basis enabling clarifying the notion of "patent-eligibility" and better understanding the "substantial more, ≫" relation[5.c]) – and models them all mathematically, except the ≫-relation, of which here only its structural characteristic is tentatively put forward.

Applying this mathematical scrutiny of D.1-D.14 – to the *KSR/Bilski/Mayo/Biosig/Alice* framework – is really awarding, as shown by the many fundamental and unquestionable new insights thus gained into the emerging areas "Innovation-R&D" and "Innovation Mathematics". It is totally unlikely they ever were achieved by just reasoning in natural legal language about therefore poorly defined SPL issues.

.b This equality insinuating sentence makes sense only on the level of notional resolution, where the BED-inCs are seen as meaningful. If this level of abstraction is somewhat reduced – for seeing how their mirror predicates on some appropriately defined spaces would define the BED-

D.10 "ABSTRACT IDEA" by *Bilski*: TT0 is called an "abstract idea" iff $\exists TT0' \neq TT0$ passing the FSTP-Test: $scope(TT0') \cap scope(TT0) \neq \Phi \wedge \exists k \epsilon [1,K] \not\ni k' : (s^{'k} > s^{k}) \wedge (s^{k'} = ip)$[5.d)].

This definition of the abstract idea property of a TT0 tells that it is patent-noneligible as preemptive, its preemptivity yet is not caused by a natural phenomenon. Thus, despite of much phony adverse rumor about the Supreme Court's notion "abstract idea" of a TT0: Its meaning is absolutely clear and very reasonable!

D.11: Induced by *Alice*, let for an ip-TT0 the term "transformation-warranting, tw" denote a category of its ip-CI-element/s' properties, modeled by "tw-inC/s" tying its ip-inC/s into a user-application, so transforming this ip-TT0 into patent-eligibility[6.a)].

Three currently broadly discussed tw-inCs – though not so termed:

1'.) In *Alice*, its specification discloses for its "transaction settling" TT0 (see **2.)**) its CI-elements and ip-inCs, but none and no combination of them is "tw-inC impacted".

2'.) *DDR*'s "customer contact" CI-element has a "look&feel" abstract idea ip-inC and the – by *Alice* and accordingly by the CAFC – "customer retention" tw-inC [160].

3'.) CAFC's recent *Myriad* decision strangely ignored, of its TT0 the CI-element "cancer indicator" – in its claims' wordings even explicitly quoted [163,183] – and its tw-inC, which makes TS(tw-inC) contain solely "BRCA1" and/or "BRCA2"[5.c)].

D.12 "PATENT-ELIGIBLE" by *Alice*: An ip-TT0 is called "patent-eligible" iff $\exists \{k^*\} \subset [1,K] : \wedge^{\forall k \epsilon [1,K]}$BED-crC0k$\gg \wedge^{\forall k \epsilon [1,K] \setminus \{k^*\}}$BED-crC0k, whereby the "$\gg$" has the meaning "{k*} transforms the latter conjunction into a user-application"[6.a)].

This definition of patent-eligibility might mislead to considering the CAFC's 2014 *DDR* decision as legally erroneous, by arguing its TT0's "customer retention" inC is in truth an ip-inC[6.a)]. But this were a legal error as, for the posc, the *DDR* specification discloses no such increase of the size of its "customer retention" domain[6.c)].

Finally, for the relation "$\gg$" just used – introduced by *Alice*, initial clarifications gained by this paper and [150, 151,153,175,171] – a comment is in place: For achieving

---

ip-inCs and their improvements – one sees that both kinds of preemptivity generating inCs are unequal: In both cases the BED-ip-inCs' mirror predicates are defined on at least 2-dimensional spaces, 1 dimension thereof in both cases representing the respective TS(ip-inC)s components of their natural laws respectively abstract ideas.

But, the mirror predicates of a natural law property modeling and of an abstract idea ip-inC are of quite different semiotics. The former semiotics tells: Nobody is capable of today explaining the internal logic of its TS(ip-inC), i.e. it got to be understood, today, as an axiom, the reasonability of which is supported by nothing else but experience. For the latter semiotics, the internal logic of its TS(ip-inC) is clearly definable by known impacts on its ip-inC of its original TS(inC) component(s), thus enabling the inventor to reasonably justify the choices he/she takes as to this ip-inC determining its CI; this ip-inC models/represents nothing today still unknown or metaphysical[6.a)].

.c The same applies for the distinction between the mirror predicates of the ip-inCs and tw-inCs. The examples 1.)/2.) for ip-inCs and 1'.)-3'.) for tw-inCs cannot disclose the whole notional complexity embodied by the relations of these notions to the above quoted Supreme Court notions, i.e. unavoidably embodied by the innovation business with ET CIs – more completely presented by [191,182].

But this complexity should not be misinterpreted as indicating that the approach to mathematically modeling a clean and operational framework for broadly consensual SPL decisions on ET CIs – based on the very abstract and nonoperational *Mayo/Biosig/Alice* framework serving the same purpose – is just a far cry. Those familiar with science/technology developments know: This approach here presents itself as already operating on intellectually firm ground – on which we got to get ahead quite a distance.

.d As shown in **1.)**, for a TT0 one of its BED-inCs may be an abstract idea, already making it an ip-TT0 (unless compensated by a tw-inC of TT0, as explained below), but also if none of its BED-inCs is an abstract idea, its TT0 nevertheless may be one by D.**10**. One might assume, TT0's preemptivity is always avoidable by adjusting the TS(BED-inC) of TT0 appropriately. This may work for some TT0s, yet in both cases there are TT0s for which this is impossible (in *Alice* reducing its TS(ip-inC)s is simply not possible, in the other case reducing TS(inC)s such that $scope(TT0') \cap scope(TT0) = \Phi$ destroys the "$\gg$", see below).

[6] **.a** The SPL semiotics of the term "X is a user-application" is: "X provides its service directly to a user" (X representing a TT0, its CI-element/s, or its/their inCs), which diametrically opposes the SPL semiotics of the term "X is a downstream-located-application" being: "X provides its service not directly to a user".

As a consequence, for a TT0, the SPL semiotics of the term "{k*} transforms the latter conjunction into a user-application" defines its meaning to be "{k*} has a patent-eligibility generating effect for the user-application by neutralizing the preemptivity generating effect of its latter conjunction's ip-inCs by disabling them from preempting other TT0s, i.e. making them defined for this user-application only[5.b)].

Accordingly – and this insight is indispensable for understanding *Bilski/Mayo/Alice* –

- an ip-inC models – as a property of the service its CI-element provides – a service of a downstream-located-application of this property and hence is necessarily preemptive, no matter whether it is of a natural phenomenon or an abstract idea sub-category of the ip-inC category, while
- a tw-inC models – also as a property of the service its CI-element provides – a service of a user-application of this property and hence cannot be preemptive, which enables it to bar this TT0/user-application from preempting other TT0s'/user- or downstream applications'.

.b The principles of both inC main-categories[6.a)], ip and tw – being noneligibility representing resp. excluding by "overriding" it[5.b)6.a)] **(!!)** – are induced by above Supreme Court's decisions as to the semiotics of refining SPL precedents for catering all parties interested in ET CIs, as *Mayo* requires. FSTP-Technology – originally designed for scientizing only the "obviousness" notion, as the BGH "*Gegenstandsträger*" decision indicated, preceding *KSR*, both reaching far into metaphysics [6[4),7[4)] – supports both inC categories [161]. While their semiotics are currently vividly discussed in smart but conventional legalese [195] – e.g. distinguishing between technical and non-technical TT0s – the ip/tw-inCs models avoid this distinction (as indefinable) and strive for more uniformity, as seemingly also being an *Alice* objective.

.c By its elements' "combinations", *Alice* allows ip- and/or tw-inCs to be BOD- or BAD- or BED-inCs.

.d In response to emerging customer requirements, TT0 improvements may lead to increasing the size of the domain of this inC and its TS – the latter from currently having the above single domain-element "customer retention" only (which by the *DDR* specification is defined as "not forwarding the customer to the Internet server of a supplier of a product if the customer clicks on this product") – such as enabling TS(inC) to comprise also values like "keep customer id secret", "keep all supplier ids secret", ….. But alike is possible with any ET CI – hence the resulting ET CIs then are considered to be separate [137]. **(!)**

.e see [7,64] – also emphasizing that it is unclear whether a pathological $CI^{FFOLLIN}$ exists, at all.

.f see [175]

consistency in SPL precedents, a threshold common to all CIs is indispensable (at least ET sub-category wise[6.b)]), and the least restrictive one – i.e. the one with maximal scope(ip/tw-CI) – is assumed to reflect the intention of the Supreme Court's above quoted line of decisions[6.a)3.c)].

This assumption is supported by the expectation, that courts would consider later simple limitations of the TS of this ip-inC as not disclosing a nonobvious CI – without explicitly explaining the difficulties to intellectually overcome for rationally arriving at them and the advantages embodied by them, especially as this broadening and/or shrinking of the scope(CI) has been anticipated by the scope(ip/tw-CI)[6.a)]. I.e.: The latter "pseudo-anticipation" would surely act as an "innovation catalyzer" – though this thinking requires refinement and feedback from the public.

D.13: For an ip/tw-CI, let "scope(ip/tw-CI)" be the modification of the S^R of the original CI as it results from the modifications of the domains and TSes of this original CI's inCs, first by its ip-inCs, making this ip-CI preemptive, and then by its tw-inCs, making this ip/tw-CI a nonpreemptive user-application.

D.14: Let the meaning of the relation "substantially more than, ≫" between an ip/tw-CI and its ip-CI be: "The ip/tw-CI's tw-inC(s) eliminate the preemptivity created by its ip-inC(s) by modifying their domains and/or TSes such that any ip-inC is defined only for and this ip-CI is transformed into a user-application ip/tw-CI of its tw/ip-inC(s)".

THEOREM: Any non-pathologic[6.d)] ET-CI may be upgraded – by using the FSTP-Test – to unassailable patent-eligibility & patentability & nonobviousness[6.e)].

Depending on the creativity effort invested in what parts and to what extent, the scope(ET-CI) would thereby vastly controllably shrink resp. grow [136].

The preceding groundbreaking definitions, consequences, and the theorem provide hitherto unavailable scientific insights into the being of an invention/innovation[5.a)] – in principle, since the 70s known to be precisely describable by (inventive) concepts, as now required to be used by *Mayo/Alice* – which is legally protectable by some FFOL Legal Innovation Norm (e.g. patent/copyright/trademark laws, institution/company regulations, business secrets, …). They represent an unexpected scientification of all kinds of IPRs in all kinds of innovations – which consequentially already enabled to design and prototype a cutting edge innovation expert system (IES), opening extremely promising perspectives at all kinds of innovation business.

In total: There is no "End of the pro-patent era", as insinuated by some [196]. The contrary is true: This era just got absolutely future-proof – world-wide.

4 A REMAINING BIG QUESTION – BROADLY IGNORED, HENCE BRIEFLY REMINDED

This inconvenient question is neither technical, nor legal, nor political – it is purely sociological and will hit soon and hard.

Already footnote 4 of [6,7] postulated that and why FSTP-Technology – the herald of Innovation Science [182] – is a new exact technology/science located on top of elementary Mathematics and below Physics, which enables a groundbreaking type of Innovation Expert Systems, IESes [161]. Just as the motorization of physical vehicles, in the first half of the 20th century, rapidly enabled a broad mechanization of all kinds of transport activities, the computerization of intellectual vehicles will rapidly enable a broad scientification of all kinds of intellectual property rights/transportation activities. In particular, IESes will massively impact on the professional activities of in particular patent experts of any kind – though much more dramatically than motorization/ mechanization has ever achieved in the world of physical transport.

The above inconvenient question then is, how this high flying prognosis for IESes – their inevitable impacts on PTOs is outlined in [163] – fits into today's professional reality, as it presents itself at such overall extremely important events like the USPTO's "Patent Quality Summit" in DC on 25.-26.03.2015.

Of this question's many facets, touched at this event, here only the most important yet vastly ignored one is addressed. It reflects that the range of patent quality issues, the USPTO must care for, is so broad that its activities have difficulties of finding a common denominator. And the same applies to the development activities of the professional profiles of its customer communities, represented by the attendees.

This most advanced kind of "digital divide" plaguing both camps, the USPTO and its customer communities, became apparent already during its first hours. The "quality related aspects of Substantive Patent Law and its currently very fast and very fundamental developments" minded participants felt evidently somewhat lost among the vast majority of participants focused on "quality aspects of the current patent eco system as it is", i.e. abstracting from such SPL developments.

The excellent main panel of this event clearly recognized this broad spread in understanding the current situation[7)]. In particular [192] addressed both these main streams with a strong bias towards the latter one – as seemingly expected by the audience and implicitly confirming its dominating "practitioner belief" that an abrupt demand of substantial increase in professional qualification is just not thinkable.

This sharply contrasts to the question written in huge letters on the wall and that the same persons [192] vividly emphasized elsewhere: How to disseminate, from mainstream one, the many advantages coming along with so seen ET CIs – by mainstream one (to be) derived from the Supreme Court's above line of decisions, i.e. this substantial increase in

professional qualification enabling to professionally leverage on them – to the in total hundred thousands of individuals of mainstream two[7]?

The majority's reluctance to notice this dissemination problem of increased IPR know-how is the reluctance to notice that the society's wealth is increasingly depending on the economical successes of ET CIs resp. of their industries, i.e. that hence the "patent eco system" must undergo a transition from its today's pre-industrial manufacture orientation to a scientized knowledge industry – as it similarly occurred previously in agriculture, clothing, food, construction, automobile, … eco systems, always generating losers and winners, indispensable for preserving the society's wealth.

The decisive distinction to such earlier transitions: ET CIs' R&D requires much higher long-term & high-risk investments than ever seen before. Due to its antiquated manufacture imprint, today's patent eco system would fail to convince investors of ET CIs' capability to guarantee a sound business model requiring such investments. By *Mayo*, the Supreme Court recognized this threat and put SPL precedents on ET CIs on the right track – namely on that of its scientification, as shown above – thus relieving it from this "pre-industrial stigma" and enabling it to gaining back investors' trust.

7 An example of this problem is the non-discussion between both main streams about the disastrous consequences that the *Mayo/Alice* decisions originally had for many PTO examiners' views about patent applications, in particular about those for ET CIs. And still today, their representative in this panel indicated in no way that the corps of examiners accept that the Supreme Court by these decisions – evidently stimulated by the economically rapidly increasing importance of ET CIs and the classical SPL precedents had proven not to be consistently applicable to them – had to pose these new intellectual challenges as to accordingly adjusting such ET CIs' Intellectual Property Rights, i.e. for optimally unfolding ET CIs' economically very beneficial potentials.

A convincing representative of a quality initiative – as to this substantially increased professional qualification – that the USPTO's first main stream is capable of establishing is its compilation and repeated refinement of the "Interim Eligibility Guidance" (IEG) and its consensus making within this main stream in its customer camp as to this all overarching SPL precedents refinement for catering ET CIs.

Yet, listening to the contributions during only the first hours of this event was sufficient to clearly recognize that there is only little common ground with the second main stream in both camps (USPTO and customer communities) as to appreciating the advantages of this significantly higher level of SPL precedents required/stimulated by the Supreme Court and now implemented by the first main stream in both camps.

This is really problematic, as the second main stream attendees at this event are the best informed representatives of this huge crowds in both camps. Their members hence must be estimated to be even more reluctant to accept that there is a challenge in their business life, which they got to master – and indeed can, as the IEG initiative shows. Though, at the expense of some unavoidable intellectual training, requiring some true efforts.

Thus, while the reception of this IEG is just the initial step to this higher level of professional qualification required in dealing with ET CIs – as shown by the CAFC's preceding experience of needed notional preciseness and the respective vogue of definitions and insights – it yet is also the most cumbersome one just as fortunately the most promising and eventually awarding one, as shown by the IES.

5 References

[1] S. Schindler: "US Highest Courts' Patent Precedents in Mayo/Myriad/ CLS/Ultramercial/LBC: 'Inventive Concepts' Accepted – 'Abstract Ideas' Next? Patenting Emerging Tech. Inventions Now without Intricacies"[*)].

[2] AIT, "Advanced Information Tech.", denotes cutting edge IT areas, e.g. Techniques of Artificial Intelligence/ Knowledge Representation/Description Logic/Natural Language& Semantics &Semiotics/System Design/….

[3] R. Brachmann, H. Levesque "Knowledge Representation & Reasoning", Elsevier, 2004.

[4] "The Description Logic Handbook", Cambridge UP, 2010.

[5] S. Schindler: "Math. Model. Substant. Patent Law (SPL) Top-Down vs. Bottom-Up", Yokohama, JURISIN 2013[*)].

[6] S. Schindler, **"FSTP" pat. appl.**: "THE FSTP EXPERT SYSTEM", 2012[*)].

[7] S. Schindler, **"DS" pat. appl.**: "AN INNOVATION EXPERT SYSTEM, IES, & ITS PTR-DS", 2013[*)].

[8] S. Schindler, J. Schulze: "Technical Report #1 on the '902 PTR", 2014[*)(soon)]

[9] S. Schindler: "Patent Business – Before Shake-up", 2013[*)]

[10] SSBG's AB to CAFC in LBC, 2013[*)].

[11] S. Schindler, **"inC" pat. appl.**: "inC ENABLED SEMI-AUTO. TESTS OF PATENTS", 2013[*)].

[12] C. Correa: "Handbook on Protection of IP under WTO Rules", EE, 2010.

[13] N. Klunker: "Harmonisierungsbestr. im mat. Patentrecht", MPI, Munich, 2010.

[14] "USPTO/MPEP: "2111 Claim Interpretation; Broadest Reason. Interpretation;]" (See App. 132a-135a)[*)].

[15] S. Schindler: "KR Support for SPL Precedents", Barcelona, eKNOW-2014[*)].

[16] J. Daily, S. Kieff: "Anything under the Sun Made by Humans SPL Doctrine as Endogenous Institutions for Commercial Innovation", Stanford / GWU[*)].

[17] CAFC En banc Hearing in LBC, 12.09.2013.

[18] SSBG AB to the Supreme Court in CLS, 07.10.2013[*)].

[19] SSBG AB to the Supreme Court in WildTangent, 23.09.2013[*)].

[20] USPTO, "Intellectual Property and the US Economy: INDUSTR. IN FOCUS", 2012[*)].

[21] K. O'Malley: Keynote Address, IPO, 2013[*)].

[22] S. Schindler, "The View of an Inventor at the Grace Period", Kiev, 2013[*)].

[23] S. Schindler, "The IES and its In-C Enabled SPL Tests", Munich, 2013[*)].

[24] S. Schindler, "Two Fundamental Theorems of 'Math. Innovation Science'", Hong Kong, ECM-2013[*)].

[25] S. Schindler, A. Paschke, S. Ramakrishna, "Form. Leg. Reas. that an Inven. Satis. SPL", Bologna, JURIX-2013[*)].

[26] SSBG AB to the Supreme Court in Bilski, 06.08.2009[*)].

[27] T. Bench-Capon, F. Coenen: "Isomorphism. and Legal Knowledge Based Systems", AI&Law, 1992[*)].

[28] N. Fuchs, R. Schwitter. "Attempt to Controlled English", 1996.

[29] A. Paschke: "Rules / Logic Programming in the Web". 7. ISS, Galway, 2011.

[30] K. Ashley, V. Walker, "From Informa. Retrieval to Arg. Retrieval for Legal Cases: ..…", Bologna, JURIX-2013[*)].

[31] Hearing in Oracle vs. Google, "As to Copyrightability of the Java Platform", CAFC, 06.12.2013.

[32] S. Schindler, "A KR Based Innovation Expert System (IES) for US SPL Precedents", Phuket, ICIIM-2014[*)].

[33] S. Schindler, "Status Report about the FSTP Prototype", Hyderabad, GIPC-2014.

[34] S. Schindler, "Status Report about the FSTP Prototype", Moscow, LESI, 2014.

[35] S. Schindler, IPR-MEMO: "STL, SCL, and SPL – STL Tests seen as SCL Tests seen as SPL Tests", in prep.

[36] S. Schindler, "Boon and Bane of Inventive Concepts and Refined Claim Construction in the Supreme Court's New Patent Precedents", Berkeley, IPSC, 08.08.2014[*)].

[37] D.-M. Bey, C. Cotropia, "The Unreasonableness of the BRI Standard", AIPLA, 2009[*)].

[38] Transcript of the Hearing in TELES vs. CISCO/USPTO, CAFC, 08.01.2014[*)].

[39] Transcript of the en banc Hearing in CLS vs. ALICE, CAFC, 08.02.2013[*)].

[40] SSBG's Brief to the CAFC in case '453[*)].

[41] SSBG's Brief to the CAFC in case '902[*)].

[42] SSBG's Amicus Brief to the CAFC in case CLS, 06.12.2012[*)].

[43] S. Schindler, **"LAC" pat. appl.**: „Semi-Automatic Generation/Customization of (All) Confirmative Legal Argument Chains (LACs) in a Claimed Invention`s SPL Test, as Enabled by Its Inventive Concepts", 2014[*)].

[44] R. Rader: "Patent on Life Sciences", Berlin, LESI, 2012.

[45] SSBG's AB to the Supreme Court as to the CII Question, 28.01. 2014[*)].

[46] S. Schindler: "Autom. Deriv. of Leg. Arg. Chains (LACs) from Arguable Subtests (ASTs) of a Claimed Invention's Test for Satisfying. SPL", U Warsaw, 24.05.2014[*)].

[47] S. Schindler: "Auto. Generation of All ASTs for an Invention's SPL Test", subm. for publ.[*)].

[48] USPTO/MPEP, "2012 ... Proc. for Subj. Matter Eligibility ... of Process Claims Involving Laws of Nature", 2012[*)].

[49] USPTO/MPEP, Supp. Examination Guidelines for Determining Compliance With 35 U.S.C. 112, Federal Register / Vol. 76, No. 27; MPEP 2171, [*)].

[50] NAUTILUS v. BIOSIG, PFC, 2013[*)].

[51] BIOSIG, Respondent, 2013[*)]

[52] Public Knowledge et al., AB, 2013[*)].

[53] Amazon et al., AB, 2013[*)].

[54] White House, FACT SHEET - ... the Presid.'s Call to Strength. Our Patent System and Foster Innovation, 2014[*)].

[55] USPTO: see home page.

[56] IPO: see home page.

[57] M. Adelman, R. Rader, J. Thomas: "Cases and Materials on Patent Law", West AP, 2009.

[58] SSBG's Amicus Brief to the Supreme Court as to its (In)Definiteness Quest's, 03.03, 2014*).

[59] S. Schindler, **"UI" pat. appl.**: "An IES Capable of Semi-Auto. Generating/Invoking All Legal Argument Chains (LACs) in the SPL Test of a Claimed Invention (CI), as Enabled by Its Inventive Concepts (inCs)", 2014*).

[60] S. Schindler: "Automatic Derivation of All Argument Chains Legally Defending Patenting/Patented Inventions", ISPIM, Montreal, 6.10.2014, updated version*).

[61] H. Wegner: "Indefiniteness, the Sleeping Giant in Pat. Law", www.laipla.net/hal-wegners-top-ten-patent-cases/.

[62] .a) CAFC decision on reexamination of U.S. Pat. No. 7,145,902, 21.02.2014*).

[63] .b) CAFC decision on reexamination of U.S. Pat. No 6,954,453, 04.04.2014*).

[64] B. Wegner, S. Schindler: "A Mathematical Structure Modeling Inventions", Coimbra, CICM-2014*).

[65] SSBG's Petition to the CAFC for Rehearing En Banc in the '902 case, 18.04.2014*).

[66] CAFC: VEDERI vs. GOOGLE decision, 14.03.2014

[67] CAFC: THERASENSE vs. BECTON & BAYER decision, 25.05.2011

[68] B. Fiacco: Amicus Brief to the CAFC in VERSATA v. SAP&USPTO, 24.03.14*).

[69] Transcript of the oral argument in U.S. Supreme Court, Alice Corp. v. CLS Bank, Case 13-298, March 31, 2014*).

[70] R. Rader, Keynote Speech: "Patent Law and Litigation Abuse", ED Tex Bench and Bar Conf., 01.11.2013*).

[71] S. Schindler, Keynote Speech: "eKnowledge of Substantive Patent Law (SPL) – Trail Blazer into the Innovation Age", Barcelona, eKNOW-2014*).

[72] .a) S. Schindler: "The Supreme Court's 'SPL Initiative': Scientizing Its SPL Interpretation Removes 3 Evergreen SPL Obscurities", Press Release, 08.04.2014*).
.b) S. Schindler: "The Supreme Court's 'SPL Initiative': Scientizing Its SPL Interpretation Removes 3 Evergreen SPL Obscurities – and Enables Automation in a CI's SPL Tests and Argument Chains", Honolulu, IAM2014S, 18.07.14*).

[73] .a) USPTO/MPEP: "2014 Procedure For Subject Matter Eligibility Analysis Of Claims Reciting Or Involving Laws Of Nature/Natural Principles, Natural Phenomena, And/Or Natural Products", [48,49], 2014*).
.b) MEMORANDUM: "Prelim. Examin. Instructions in view of *Alice v. CLS*", 25.06.2014*).

[74] B. Wegner: "The Mathematical Background of Proving an Inventive Concepts Based Claimed Invention Satisfies SPL", 7. GIPC, Mumbai, 16.01.2015.*)

[75] CAFC Order as to denial [65], 27.05.2014

[76] D. Crouch: "En Banc Federal Circuit Panel Changes the Law of Claim Construction", 13.07.2005*).

[77] Video of the USPTO Hearing, 09.05.2014*).

[78] R. Rader, Keynote Speech at GTIF, Geneva, 2014 and LESI, Moscow, 2014

[79] S. Schindler: "On the BRI-Schism in the US National Patent System …", publ. 22.05.2014.*)

[80] SSBG's Pet. for Writ of Cert. to the Supr. Court in the '902 case, Draft_V.133_of_[121], 14.07.2014*).

[81] S. Schindler: "To Whom is Interested in the Supreme Court's Biosig Decision", 04.06.2014*)

[82] R. DeBerardine: "Innovation from the Corporate Perspective", FCBA, DC, 23.05.2014*).

[83] SSBG's Petition to the CAFC for Rehearing En Banc in the '453 case, 09.06.2014*).

[84] CAFC's Order as to denial [83], 14.07.2014*).

[85] CAFC: "At Three Decades", DC, 2012.

[86] Sigram Schindler Foundation: "Transatlantic Coop. for Growth and Security", DC, 2011.

[87] DPMA: "Recent Developments and Trends in US Patent Law", Munich, 2012.

[88] FCBA: "Innovation, Trade and Fiscal Reality", Colorado Springs, 2013.

[89] LESI: GTIF, Geneva, 2014.

[90] FCBA: "Sharpening Case Management", Asheville, North Carolina, 2014

[91] B. Wegner, S. Schindler: "A Mathematical KR Model for Refined Claim Construction II", subm. for publication.

[92] SSBG's Petition for Writ of Certiorari to the Supreme Court in the '453 case, 06.10.2014*).

[93] E. Morris: "What is 'Technology'?", IU I.N.*)

[94] E. Morris: "Alice, Artifice, and Action – and Ultramercial", IU I.N., 08.07.2014*).

[95] S. Schindler, ArAcPEP-MEMO: "Artifice, Action, and the Patent-Eligibility Problem", in prep., 2014.

[96] A. Chopra: "Deer in the Headlights. Response of Incumbent Firms to … ", School of Management, Fribourg, 2014*).

[97] S. Schindler, DisInTech-MEMO: "R&D on Pat. Tech.: Efficiency and Safety Boosting", in prep., 2014.

[98] G. Boolos, J. Burgess, R. Jeffrey: "Computability and Logic", Cambridge UP, 2007.

[99] A. Hirshfeld, Alexandria, PTO, 22.07.2014*).

[100] C. Chun: "PTO's Scrutiny on Software Patents Paying Off", Law360, N.Y., 22.07.2014*).

[101] P. Michel, Keynote, Alexandria, PTO, 22.07.2014.

[102] D. Jones, Alexandria, PTO, 22.07.2014.

[103] R. Gomulkiewicz, Seattle, CASRIP, 25.07.14.

[104] M. Lemley, Seattle, CASRIP, 25.07.2014.

[105] D. Jones, Seattle, CASRIP, 25.07.2014.

[106] B. LaMarca, Seattle, CASRIP, 25.07.2014.

[107] J. Duffy, Seattle, CASRIP, 25.07.2014.

[108] J. Pagenberg, Seattle, CASRIP, 25.07.2014.

[109] M. Adelman, Seattle, CASRIP, 25.07.2014.

[110] B. Stoll, Seattle, CASRIP, 25.07.2014.

[111] R. Rader, Seattle, CASRIP, 25.07.2014.

[112] E. Bowen, C. Yates: "Justices Should Back Off Patent Eligibility, …", L360, 25.07.2014*).

[113] S. Schindler: "The CAFC's Rebellion is Over – The Supreme Court, by *Mayo/Biosig/Alice*, Provides Clear Guidance as to Patenting Emerging Technology Inventions", published 07.08.2014*).

[114] S. Elliott: "The USPTO Patent Subj. Matter Eligi. Guidance TRIPSs", 30.07.2014*).

[115] W. Zheng: "Exhausting Patents", Berkeley, IPSC, 08.08.2014*).

[116] R. Merges: "Independent Invention: A Limited Defense of Absolute Infringement Liability in Patent Law", Berkeley, IPSC, 08.08.2014*).

[117] J. Sarnoff, Berkeley, IPSC, 08.08.2014.

[118] H. Surden: "Principles of Problematic Patents", Berkeley, IPSC, 08.08.2014*).

[119] www.ssit.de/2013/33/multiple-sklerose-medikament-tecfidera/seite-2*).

[120] J. Merkley, M. Warner, M. Begich, M. Heinrich, T. Udal: "Letter to Hon. Penny Pritzker", DC, 06.08.2014*).

[121] SSBG's Petition for Writ of Certiorari to the Supreme Court in the '902 case, 25.08.2014*).

[122] D. Parnas, see Wikipedia.

[123] E. Dijkstra, see Wikipedia.

[124] S. Schindler: "Computer Organization III", 3. Semester Class in Comp. Sc., TUB, 1974-1984.

[125] S. Schindler: "Nonsequential Algorithms", 4. Semester Class in Comp. Sc., TUB, 1978-1984.

[126] S. Schindler: "Optimal Satellite Orbit Transfers", PhD Thesis, TUB, 1971.

[128] R. Feldman: "Coming of Age for the Federal Circuit", The Green Bag 2014, UC Hastings.

[129] G. Quinn: "Judge Michel says *Alice* Decision 'will create total chaos'", IPWatchdog, 06.08.2014*).

[130] G. Frege: "Function und Begriff", 1891.

[131] L. Wittgenstein: "Tractatus logico-philosophicus", 1918.

[132] B. Wegner, MEMO: "About relations (V.7-final)", 25.04.2013*).

[133] B. Wegner, MEMO: "About conjunctions of predicates/concepts, scope and solution of problems", 20.08.2013.

[134] B. Wegner, MEMO: "A refined relation between domains in BADset and BEDset", 18.09.2014.

[135] H. Goddard, S. Schindler, S. Steinbrener, J. Strauss: FSTP Meeting, Berlin, 29.09.2014.

[136] S. Schindler: "Tutorial on Commonalities Between System Design and SPL Testing", sub. for pub.*).

[137] S. Schindler: "The Rationality of a Claimed Invention's (CI's) post-*Mayo* SPL Test – It Increases CI's Legal Quality and Professional Efficiency in CI's Use –Iits Semiotics Inspiing the Inventivity to/in CI's Further Development", in prep.

[138] S. Schindler: "The Supreme Court's Guidance to Robust ET CI Patents", ICLPT, Bangkok, 22.01.2015*).

[139] Supreme Court's Order as to denial [121], 14.10.2014*).

[140] S. Schindler: "§ 101 Bashing or § 101 Clarification", published 27.10.2014*).

[141] BGH, "Demonstrationsschrank" decision*).

[142] B. Wegner, S. Schindler: "A Mathematical KR Model for Refined Claim Construction II", submitted for pub.

[143] … Press, …… to go into [137]……

[144] "Turmoil …..", see program of AIPLA meeting, DC, 23.10.2014

[145] "Dark side of Innovation", …… to go into [137]……

[146] D. Kappos: About his recent west coast meetings, AIPLA, DC, 23.10.2014.

[147] Transcript of the CAFC Hearing in *Biosig* case, 29.10.2014*).

[148] R. Rader: Confirming that socially inacceptable CIs as extremely preemptive, such as for example [119][2)], should be patent-eligible, AIPLA meeting, DC, 24.10.2014.

[149] A. Hirshfeld: Announcing the PTO's readiness to consider also hypothetical CIs into its resp. guideline, AIPLA meeting, DC, 24.10.2014.

[150] S. Schindler: "Alice-Tests Enable "Quantifying" Their Inventive Concepts and thus Vastly Increase the Robustness" of ET Patents – A Tutorial about this Key to Increasing a Patent's Robustness –", USPTO&GWU, 06.02.2015*), also ABSTRACT, see also [175]*).

[151] S. Schindler: "*Biosig*, Refined by *Alice*, Vastly Increases the Robustness of Patents – A Tutorial about this Key to Increasing a Patent's Robustness –"*).

[152] S. Schindler: "Automatic Derivation/Reproduction of Legal Argument Chains (LACs), Protecting Patens Against SPL Attacks", Singapore, ISPIM, 09.12.2014*).

[153] S. Schindler: "Practical Impacts of the Mayo/Alice/Biosig-Test – A Tutorial about this Key to Increasing a Patent`s Robustness", 2015 IP Scholars Roundtable, Drake University Law School, 27.03.2015*)

[154] CAFC Decision in *Interval*, 10.09. 2014*).

[155] S. Schindler: "A Tutorial into (Operating) System Design and AIT Terms/Notions on Rigorous ET CIs' Analysis by the Patent Community", in prep.

[156] CAFC Decision in *DDR*, 05.12. 2014*).

[157] USPTO: "2014 Interim Guidance on Patent Subject Matter Eligibility & Examples: Abstract Ideas", 16.12.2014*).

[158] Supreme Court's Order as to denial [92], 08.12.2014*).

[159] CAFC Decision in *Myriad*, 17.12.2014*).

[160] S. Schindler: "The Supreme Court's *Mayo/Myriad/Alice* Decisions, The PTO's Implementation by Its Interim Eligibility Guidance (IEG), The CAFC's *DDR &Myriad* Recent Decisions – Clarifications&Challenges"*), publ. 14.01.2015*), its short version*), and its PP presentation at USPTO, 21.01.2015*)..

[161] S. Schindler: "The Innovation Expert System, IES: Philosophy & Functionality &, Mathematical Foundation – A Prototype Outl.", 7. GIPC, Mumbai, 16.01.2015*).

[162] CAFC Decision in *CET*, 23.12.2014*).

[163] S. Schindler: "The USSC's Mayo/Myriad/Alice Decisions: Their Overinterpret. vs. Oversimplification of ET CIs – Scientific. of SPL Prec. as to ET CIs in Action: The CAFC's Myriad &CET Decisions", USPTO, 07.01.2015*).

[164] J. Schulze, D. Schoenberg, L. Hunger, S. Schindler: "Introduction to the IES User Interface of the FSTP-Test ", 7. GIPC, Mumbai, 16.01.2015, PPP*).

[165] "ALICE AND PATENT DOOMSDAY IN THE NEW YEAR", IPO, 06.01.2015*).

[166] S. Schindler: "Today's Substantive Patent Law (SPL) Precedents and Its Perspectives, Driven by ET CIs", 7. GIPC, Mumbai, 15.01.2015*).

[167] R. Sachs: "A Survey of Patent Invalidations since *Alice*". Fenwick & West LLP, Law360, New York, 13.01.2015*).

[168] S. Schindler: "PTO's IEG Forum – Some Aftermath", publ. 10.02.2015*).

[169] Agenda of this Forum on [157], Alexandria, USPTO, 21.01.2015*).

[170] G. Quinn: "Patent eligib. forum discuss. examiners application of *Mayo/Myriad/Alice*", IPWatchdog, 21.01.2015*)

[171] S. Schindler: "Semiotic Impacts of the Supreme Court's Mayo/Biosig/Alice Decisions on Legally Analyzing Emerging Technology Claimed Inventions (ET CIs)", 16th Int. Roundtable on Semiotics of Law, Hilo, 29.04.2015, draft sub. f. pub...

[172] USSC Decision in *Teva*, 20.01.2015*).

[173] USSC Decision in *Pullman-Standard*, 27.04.1982*).

[174] USSC Decision in *Markman*, 23.04.1996*).

[175] S. Schindler: "Increasing a Patent's Robustness by 'Double Quantifying' Its Inventive Concept as Implied by *Mayo/Alice*", WIPIP. USPTO&GWU, 06.02.2015*).

[176] R. Rader: Questions as to the FSTP-Test, WIPIP, USPTO&GWU, 06.02.2015.

[177] D. Karshtedt: "The Completeness Requirement in Patent Law", WIPIP, USPTO&GWU, 06.02.2015*).

[178] O. Livak: "The Unresolved Ambiguity of Patent Claims", WIPIP, USPTO&GWU, 06.02.2015*).

[179] J. Miller: "Reasonable Certain Notice", WIPIP, USPTO&GWU, 06.02.2015*)

[180] S. Ghosh: "Demarcating Nature After Myriad", WIPIP, USPTO&GWU, 06.02.2015*)

[181] CAFC Decision in *Cuozzo*, 04.02.2015*).

[182] S. Schindler: "Patent/Innovation Technology and Science", Textbook, in preparation.

[183] S. Schindler: "The Mayo/Alice SPL Terms/Notions in FSTP-Technology & PTO Initiatives", USPTO, 16.03.2015*).

[184] S. Schindler: "PTOs Efficiency Increase by the FSTP-Test, e.g. EPO and USPTO", LESI, Brussels, 10.04.2015..

[185] R. Chen: Talking politely of "tensions" in stating a CI's indefiniteness by the BRI, PTO/IPO-EF Day, 10.03.2015.

[186] A. Hirshfeld: Reporting about the PTO's view at the progress of the IEG work, PTO/IPO-EF Day, 10.03.2015.

[187] P. Michel: Moderating a panel about the SPL paradigm refinement by *Mayo/Alice*, PTO/IPO-EF Day, 10.03.2015.

[188] P. Michel: Asking this panel as to dissemination of *Mayo/Alice* understanding, PTO/IPO-EF Day, 10.03.2015.

[189] M. Lee: Luncheon Keynote Speech, PTO/IPO-EF Day, 10.03.2015*).

[190] A. Hirshfeld: Remark on EPQI's refinement of patent application examination, PTO/IPO-EF Day, 10.03.2015.

[191] B. Wegner, S. Schindler: "A Mathematical KR Model for Refined Claim Construction III", in prep.

[192] M. Schecter, D. Crouch, P. Michel: Panel Discussion, Patent Quality Summit, USPTO, 25.03.2015.

[193] Finnegan: 3 fundamental current uncertainties about SPL precedents, Patent Quality Summit, USPTO, 25.03.2015.

[194] S. Schindler: "post-*Mayo/Biosig/Alice*: The Precise Meanings of SPL Terms for ET CIs", this paper, publ.08.04.2015*).

[195] R. Stoll: "Federal Circuit Cases to Watch on Software Patentability – Planet Blue", Patently-O, 06.04.2015*).

[196] See the resp. prominent panel at the IPBCGlobal'2015, San Francisco, 14-16.06.2015*).

*) available at www.fstp-expert-system.com

"Kids' Friendly Factor" in Urban Spaces

[1] **Anahita Mohammadi,** [2]**Ali Jabbari Jahromi,** [3]**Azadeh Alighanbari**
[1]Architecture Department, IAU University of Beyza, Iran
[2]Architecture Department, IAU University of Shiraz, Iran
[3]Architecture Department, IAU University of Shiraz, Iran

Abstract- *Urban spaces as it is understood from their names are public spaces for the use of different racial, gender, age and etc. as long as human definition is applicable on an existence, its right of using urban spaces is undoubtable. What is going to be highlighted in this article is going to analyze users in kid's age group in categorizing according to their age.*

Urban spaces often adjust their informant parameters with limitations, needs and potentials of users of various age groups to improve the quality of their performance, parameters such as availability, circulation, depth of space and etc.

Since urban spaces are most obvious peripheral display of our existence and their role is to continuously interpret and transfer hidden and exposed codes risen dominant culture and being valued by it. When we talk about urban spaces and those hidden codes for kids, in fact we are talking about discrete species of intelligent creatures living among us.

This is a language which is unknown and not understandable for kids because of the lack of mutual intellectual /empirical background in them. Despite the fact that in statistical analyzes, children had always formed one of the considerable population groups in the range of urban spaces' users, functionally they have noticeably minor effect on parameters consisting urban spaces.

This article effort to define the first steps in the improvement of existing urban spaces for kids. What will be proposed ultimately is a try to reach to a standard to define and evaluate "kid's friendly factor" for an urban space.

In this way we would be able to analyze the effective parameters on improving the kids' way of using and involving in urban spaces. We also move forward towards definition of the next steps in converting their aspect of effect into measurable numerical components to be proceeded.

This analysis will proceed to a proposal to evaluate existing urban spaces based on optimum quality usage for kids.

The predictable practical step will be the manipulation and adjustment of these effective parameters, in a way that we add to kids' portion of their undoubtable/undeniable right for using urban spaces.

Keywords- Urban space, Kids, Accessibility, Kid's friendly, Environment, Parameter

1 Introduction- Kids' share of Urban Space

Despite the fact that seemingly, kids have the most joyful population breakdown among the population groups of urban space users, in fact it is their high adjustability that makes them so happy.

Other population groups of users consciously seek fulfilling their needs and expectations in urban spaces and expect the adjustment of urban spaces with a noticeable voice using every and each possible media. But kids confront this subject in urban spaces in totally different way. They adjust the way of fulfillment of their needs with situations in a unique way and define and fulfill their mental and physical needs in the framework of available facilities.

In an obvious contradiction, this high level of adjustability provokes urban designers not to

consider the needs of these users' age group as a basic parameter in their priorities.

Urban spaces in their general concept are either formed organically or dictated to face of cities as infra-structures by urban designers, city authorities and etc. In that process while urban spaces conveyed a general concept for adults in all their aspects, these where playgrounds which basically dedicated to kids as their urban space. The problem is that playgrounds even in small scales, that is to say as a part of urban spaces also are considered as luxury elements in urban spaces' designation. In this situation evaluation of "kids' friendly factor" for existing urban will be one of the shortest ways for this age group's maximum partnership.

This article will follow the possibility of empowering these parameters in the mentioned spaces by analyzing the effective parameters considered in defining "kid's friendly factor" in urban spaces.

Figure 1- Kids' Limitless Creativity

2 Adjustability versus the quality of utilization

Children have a unique ability to adjust with the conditions surrounding them and adjust their needs with the existing/available conditions. This fact has led to the difficulty of analysis and perception of the feedbacks based on each individual's experiences. In this age group in the case of perception, it would be along with vague and haywire results. Creative and wild imagination of kids alters almost all physical environments to a personal adventure scene without considering true spatial essential requirements for these none demanding users. Therefore these should be humanism researchers in the fields related to environmental factor affecting kids' social behaviors are the ones of the reliable resources for cognition of the essential parameters for this subject.

Figure 2- Kids Accommodating With Their Environment

3 Parameters affecting quality of usage of urban spaces for kids

As it is seen in the studies, in many of the cases the factor that makes an urban space to have high quality of performance in an adult designer/user's point of view can actually be in a direct conflict with the kid user's utilization quality of the same environment. Because of that, the mentioned values in these parameters have been analyzed from special viewpoint diverted to the kid user's age group.

The parameters being discussed naturally have a relative nature and are under the impact of numerous environmental/cultural factors. Since urban design itself, is naturally an interdisciplinary field, the valuing criterion of these parameters is mostly recognized based on successful and long term experiments. This fact exposes relative nature of the parameters being discussed.

What we seek ultimately is creating a process-based structure in evaluation of "kids' friendly factor" of an existing urban space based on considering and validating of these parameters.

3.1 Accessibility

Accessibility in its general concept is considered as one of the effective factors in formation and resurrection of an urban space. In fact a high percentage of urban spaces are practically defined through intersections of pedestrian and vehicles' pathways. As it was

mentioned, although high level of accessibility leads to the various kinds of users and as a result to enriching urban spaces with intergroup relations, on one hand it reduces the quality of kids' utilization of the spaces being discussed to a considerable extent

3.1.1 Vehicles' Accessibility

Vehicles' accessibility for a range of adult users means the possibility of using urban spaces with economizing of displacement time in an optimum period. This may be discussed for another range of adults as a factor decreasing the utilization quality of urban spaces with consideration of aesthetical parameters and also their impact on disorder level in silence of the environment.

This happens for kids in a completely different framework.

The sheer vicinity of vehicles' accessibility to children's activity area is considered as an undesirable element. At first the most obvious designable instance is the preservation of kids' physical security and health which is posed as the first premiership. From this viewpoint any vicinity to the direct access of urban area to the vehicles' accessibility is considered as a negative point. Because if an urban area doesn't have the ability to secure the child's physical health, any more analyzes for kids' desirable quality parameters will be redundant.

On the other hand, the effectiveness of vehicles' accessibility will weaken the active informational stream between the kid and urban space by reducing the amount of kids' perception of environmental feedbacks from urban spaces because of the high attraction of its dynamic graphical attraction and also overshadowing active sounds of urban spaces including other kids, parents, natural elements such as wind and etc.

Figure 3- Vehicles as Potential Hazards

3.1.2 Pedestrians/ Internal Accessibility

Maybe it can be stated that actually an urban space has been founded based on the construction of a set of pedestrian's accessibilities and is settled by that .The term "urban space" will be meaningless without pedestrian's accessibility. Since our concentrations is on the way these accessibilities affect children's activity in urban spaces it can be said that the existence of open circles of pedestrian accessibility in an urban environment can be the best option for enhancing the quality of kids' utilization of an urban space. Closed circles of accessibility at one hand create a suitable environment for kids' less interfered activity (by other age groups) and on the other hand don't affect kids' connection with their intermediates (parents for instance) at the same time.

3.2 Depth of perceptible space/visual range

Depth of space is one of the parameters that valuation of that in an urban space has a straight connection with age group of the users under study.

We define "visual range" in any specific location of a as a physical/mental perceptible range by the user in that point of urban space, If that point likewise has the capability of nomination as a part of the urban space.

Sight sense works as the most important and essential intermediate and is a tool for deep perception of surrounding environment by kids.

For adult users the depth of perceptible space is defined with respect to the valued physical/mental elements by the user himself and according to his experimental history of that space. Elements such as high rise urban elements, meaningful graphics and ads or

occupancies with high level of priority (economical, functional and etc.) in the range of that urban area under study.

For kids, because of the lack of experimental and intellectual precedent, the abovementioned parameters including cultural codes are substituted with dynamic/statics graphical attractions to define dimensions of perceptible space's depth.

We have to consider that uneven pace in process of completion of kids' visual ability development and sight sense as they grow, changes these dimensions into a complete relative factor even for kids of not the same age.

Possessing minimum desirable dimensions of space depth for kids in an urban space is considered as a factor with positive value in kids' optimal usage of that urban space.

3.3 Diversity in Height

In design process/ analyzing, diversity in height at some degree considered as a positive element for an urban space .Diversity in height will leads to a creation of a desirable hierarchy meant by urban designer/planner by detaching different zones of an urban space base on the occupancy, categories of users and etc.

For kids, this factor however would be analyze in a different framework as well and has comes with noticeable limitations and regulations for kids.

If this kind of diversity prevent an urban space to provide the previous mentioned of spatial depth for kids it would be valued as a negative impact on the quality of utilizing of that urban space for kids. This parameter could also work in the other way as well.

For instance and generally speaking, if that mentioned diversity in an urban space is provided by a sloped walkway such as a ramp, it will create designers desired diversity and keeps/provides required depth of space for optimizing the usage quality for kids as users in that urban space .

Creating proper physical security and health risks for kids are other subjects to be considered while analyzing height diversity in any urban space under study.

3.4 Color

Most of the times in process of designing/creating an urban space, color is hidden and conveyed in and by the materials in an urban space. This fact also has been accepted for adult users as well. It means for them a specific color will call out a specific material and vice versa.

For kids there is totally different story behind and individual color and/or combinations of them. For them color comes with a strong emotionally magnified impact which will brings out certain emotions in response. According to studies color would be the most significant sense to create a meaningful connection between a kid and his/her surrounding environment. For them colors are defined as independent mediums no matter which materials are used to carry them. For that we are able to implement that colors can actually be the most efficient parameter among those discussed so far.

Color can define any space for kids and translate the hidden codes within that space in a clear emotional way, codes such as security, danger, dynamism, excitement and etc. Utilizing of proper colors while considering the existing depth of space can emotionally bring out both positive and negative potentials of an urban space for kids.

For instance while green can point out a stress-free environment, red will express a dynamic one, or using yellow and black colors beside each other will point out a potential hazard in an environment.

In fact color has the ability to interpret an urban space into known feelings for kids and by that can plays the most important role to optimize the quality of usage of an urban space for kids.

Figure 4- Color Streets to Define Them

4 Analytical Calculations

This is an analytical calculation of these parameters that shows how kid friendly is an urban space.

As we mentioned before, the nature of analyzing and evaluating urban spaces is still relative even if users are categorized by age, gender etc. This nature will force us to be conservative in the process of converting mentioned parameter into numerical one and even more important than that when we are adding those positive/negative numerical values in purpose of reaching to an assumption for a value for "kids' friendly factor" in analyzing any existing urban space.

No need to mention that, these parameters are only few of many parameters to affect the subject and only a process based approach will keep the way open for new parameter to come to equation at some point in process. Also as urban spaces are considered successful or not based on their long term performance –which itself is a subject to change – a self-modified process is advised to cover the subject.

Using behavioral science researchers in a process oriented methodology ti evaluate these parameters can be the next step to recognize how "kid friendly" an existing urban space is.

5 Conclusions

Final goal would be first to adjust the existing urban spaces using their potentials to make them better environments for kid to utilize and second to emphasise on cognition an utilizing these parameters as essential parts of studies in process of design for new developments in urban spaces.

Creating a process oriented system to design urban spaces would be the ultimate goal to this approach. The structure of this system should be based on behavioral analyses of users by incorporating proper mathematics for that purpose.

6 References

[1] Abcrcrobie, C. F. Stanley , 1991, places for play , urban , open space Academy editions , London.

[2] Broto , carles , 2006 , Great kid's spaces , translated by amber ockrassa , links.

[3] Church , Josef , Steven Gorge , A psychology of the grewing person .

[4] Azimi , Sirous ,2002 , child's psychiatry , Amir Kabir .

[5] Craw , Alice W. , 1998 , child's psychiatry , translated by Hamedani , Moshfegh , Amir Kabir .

[6] Alkaiid , David , 1996 , Child's growing and education in Piazheh philosophy , translated by Naely , Hosein , Astan-eh-Ghods .

[7] Dr. Amably , T. , 1997 , Kids' creativity , translated by Dr. Ghasem-zadeh , Hasan , Azimi , Parvin , Donyayeh now , Tehran.

[8] Izadpanah Jahromi , Ayda , 2005 , Kids : city & play – process , basis and regulations of kids' playground design - , National municipalities organization .

[9] Jabbari Jahromi , Ali , Alighanbari , Azadeh 2012 , ,"sustainability of functional changes in real-time analyses of the socio-spatial structures",IKE2008,CSREA press.

[10] Mohammadi , Anahita , Khabiri , Mehdi 2012 , ," Emotional knowledge engineering: Children as our innocent opponents in urban spaces' ownership ",IKE2012,CSREA press.

SESSION

SECURITY, PRIVACY AND RELATED METHODS + IDENTITY MANAGEMENT METHODS

Chair(s)

TBA

Multi-Participant Encryption-Steganography Technique

Hamdy M. Mousa
Faculty of Computers and Information, Menoufia University, Egypt
hamdimmm@hotmail.com

Abstract: The security is very important when transmitting sensitive data through internet. The paper proposes multi-participant encryption-steganography technique (M-PEST) for images in order to make the technique more protection and less predictable. In this technique, sort the secret image pixels and divide the sorted image index into multi-participant with same size of secret image. Then, encrypt multi-participant and sorted image pixels, transpose and rotate them. After that, generate the keys vector that embedded into cover image using bit exchange method. Transmit the encrypted sorted image pixels and multi-participant indices in image format file. The keys vector is used to decrypt the secret image. Experimental results demonstrate that M-PEST technique has multilayer protection stages against different attacks and higher level of security based on the multi-sharing and difference between histograms of the secret image and multi-participant so, encrypted multi-images are acceptable. The reconstructed image is same as secret original.

Keywords: Data sorting, multi-share, cryptography, bit exchange, hiding technique.

1. Introduction

Maintaining the secrecy and confidentiality of images is very important when transmitting them over secured or unsecured channel. Cryptography is the technique of keeping private information from unauthorized access but authorized parties can decode it [1]. The oldest encrypted text or cipher text was found in ancient Egypt. The common technique that used to maintain security of image in storage and transmission over the network is encryption. But encryption is not enough for protection and steganography provides further security by hiding the secret text into a seemingly invisible image [2]. Visual cryptography is encryption technique to hide sensitive information in images. The secret sharing scheme divides the image into different shares and distributes them. The original image is revealed, when the shares are superimposed [3].

Noar and Shamir in 1994 are proposed Visual cryptography that is applied the Human visual system to decrypt the secret image without any computational and cryptographic algorithms [4]. The proposed scheme is involved breaking up the image into n shares so that only someone with all n shares could decrypt the secret image by superimposing the shares. They assume that the image is composed of black and white pixels, and each pixel is encrypted separately.

In a (2, 2) Visual cryptography scheme, the original image is divided into two shares such that each pixel is replaced with a non-overlapping block of two or four sub-pixels in the original image. The original image will be decoded if there are all shares. In this scheme the black pixels can be reconstructed perfectly but the white pixels cannot [5, 6].

The Visual cryptography scheme is a secure method that provides authentication by breaking image into shares [7, 8]. A visual cryptography scheme for a set P of n participants is a method to encode a secret image into n shadow images called shares, where each participant in P receives one share. The secret image is recovered by using certain qualified subsets of participants, but other, forbidden, sets of participants have no information on secrete image [9, 10, 11]. A k-n secrete sharing visual cryptography scheme for color image where encryption of image is done using random number generator [12]. A (2, N) visual cryptography technique can be used in banking application [13].

The visual cryptography schemes were applied to only black and white images until the year 1997. Verheul and Van Tilborg [14] is the first colored visual cryptography scheme. A colored secret image uses the concept of arcs to construct a colored visual cryptography scheme. One pixel is transformed into m sub pixels, and each sub pixel is divided into c color regions in c-colorful visual cryptography scheme. But in both of the schemes, the shares generated were meaningless.

Chang and Tsai [12] implemented color visual cryptography scheme for sharing a secret color image and also to generate the meaningful share to transmit secret color image. In the color visual cryptography for a secret color image there are two significant color images are selected as cover images which are the same size as the secret color image.

Pixel expansion and the quality of the reconstructed secret image have been a major issue of visual secret sharing schemes that maintains the perfect security and the size of the original image.

The paper introduces multi-participant Encryption-Steganography technique (M-PEST) for any images types. In this technique, the secret image pixels are sorted and the sorted image index is partitioned into participants with same size of secret image. Then, encode participants' matrices and sorted image pixels using different keys and

the encrypted keys vector is generated and embedded into host image. Transmit host image and the encrypted image pixels and participants' matrices in image format file. Experimental results prove that reconstructed image is typical copy of secret image. And they also demonstrate that proposed technique maintains the perfect security and the size of the original image. The paper is structured as follows: in section 2, M-PEST is explained, and then in section 3 experimental results and discussion and we prove that it satisfies the presented security. Finally, the last section is conclusion.

2. M-PEST Technique

In many applications, the need of security increases. Encryption is a common technique that has application in various fields includes internet communication, medical imaging and military communication.

The encryption is not sufficient so steganography is the supplementary to encryption. It is not the replacement of encryption. But, steganography beside to encryption makes sensitive data more security. In this paper, M-PEST is proposed to protect any type of gray and RGB images. The main steps of M-PEST are explained as the following.

At the beginning, read secret Image (I_1) and separate it to three components if it is RGB image. Then, convert each component to vector. For every component vector, sort its pixels and the resulted index are partitioned to three vectors using two keys or three keys according to the algorithm for making the resultant an 8 bit image in range 0 to 255. Then, encrypt four vectors (three index partitions and image pixels' component) using only one key or more. Reshape four vectors to original image size. After that, divide these 4/12 images into blocks and transposition some blocks of images and/or rotate left/right. At this moment, construct keys vector that contains the algorithm number, images arrangement, keys values, rotation and transposition. Embed the keys vector into cover image (I_2). In the embedded stage, the threshold value can be used that achieved this condition; the pixel value is greater than even threshold value dependent on the algorithm number that used. The algorithm number is embedded in predefined bits in cover image and defines the arrangement of keys vectors and the length of each key, rotation and its direction; transposition blocks number, block size and the threshold use or not and its value. The generated keys vector may be:

AlgNo, Key1, Key2, Key3, ……. , ImagesArrangement, Rot, Dir, blockSize, BlockTrans, …… , ...

In general all ranges doesn't defined or limited in M-PEST, but in our implementation, the range of AlgNo is from 1 to 10, the block size varies from 8 pixels to 50% of image size range, there are no more 5 block transposition and one time left/right shift with shifting value from 0 to 20% of Image Size range. There is an example of Keys vectors that is shown in Figure 1.

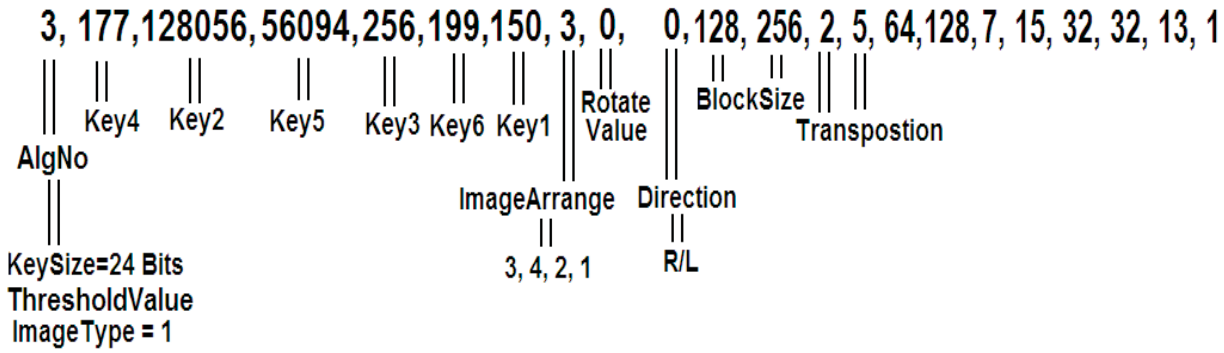


Fig. 1: keys vectors example

In the Embedded keys vector stage, read and prepare cover file and the starting pixel to embed the vector is previously defined, the encrypted keys vector is embedded into the cover file by changing the least significant bit.

The Pseudocode of Encoding and Decoding secure multi-participant Encryption-Steganography are shown in Fig. 2 and Fig. 3.

```
Input: Read Secret Image, Cover image
Output: encrypted images, Cover image
Define the algorithm number or randomly generated.
If RGB image then
    Component = 3
    Separate components
Else
    Component = 1
Endif
For each component
    Convert the secret image component to ImageVector.
    [Index, SortedImage] = sort ImageVector.
    Partition Index to 3 IndVector in range [0:255] using 2 keys.
    Encrypt 3 IndVector and SortedImage using 1 or more keys.
    Reshape 3 IndVector and SortedImage to the Secret Image size.
    Divide IndVector and SortedImage into blocks
    Transposition 0 or more blocks of the 3 IndVector and SortedImage matrices.
    Rotate 3 IndVector and SortedImage matrices or not.
    Save 3 IndVector and SortedImage matrices in 4 images files.
End for
Construct and binarize the Keys vector values.
Embed the Keys vector values into cover image.
Transmit 5/13 images files.
```

Fig. 2: Pseudocode of proposed encoding-encryption

```
Input: Read 5/13 images
Output: secret image
Extract the encrypted Keys vector from cover image.
Define Algorithm number and Keys vector values
Rearrange the 4/12 encrypted images (rotation/transposition)
Convert 4/12 encrypted images to vectors
Calculate the Index
Reconstruct the Secret image.
Save Secret image.
```

Fig. 3: Pseudocode of proposed decoding - encryption

3. Implementation and Evaluation Results

The M-PEST technique is implemented using MATLAB 2009 on Windows 7 platform and run the program with different size and type secret images and different cover images types.

In this section, some experiments are carried out to prove the efficiency of the proposed M-PEST technique. The original image, reconstructed-image and cover image are shown in Figure 4. The peak signal-to-noise ratio (PSNR) represents the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. On reconstructing the secret image is obtained back without any loss. The quality of the secret image can be analyzed by using the PSNR value of original secret image and reconstructed image. The value of PSNR is infinity, meaning that the two images are identical. The PSNR value of cover image before/after embedding keys vector is greater than 50 dB. The twelve encrypted images are shown in Figure 5.

Secret Image	Reconstructed Secret Image	Cover Image

Fig. 4: The original secret images and encrypted images

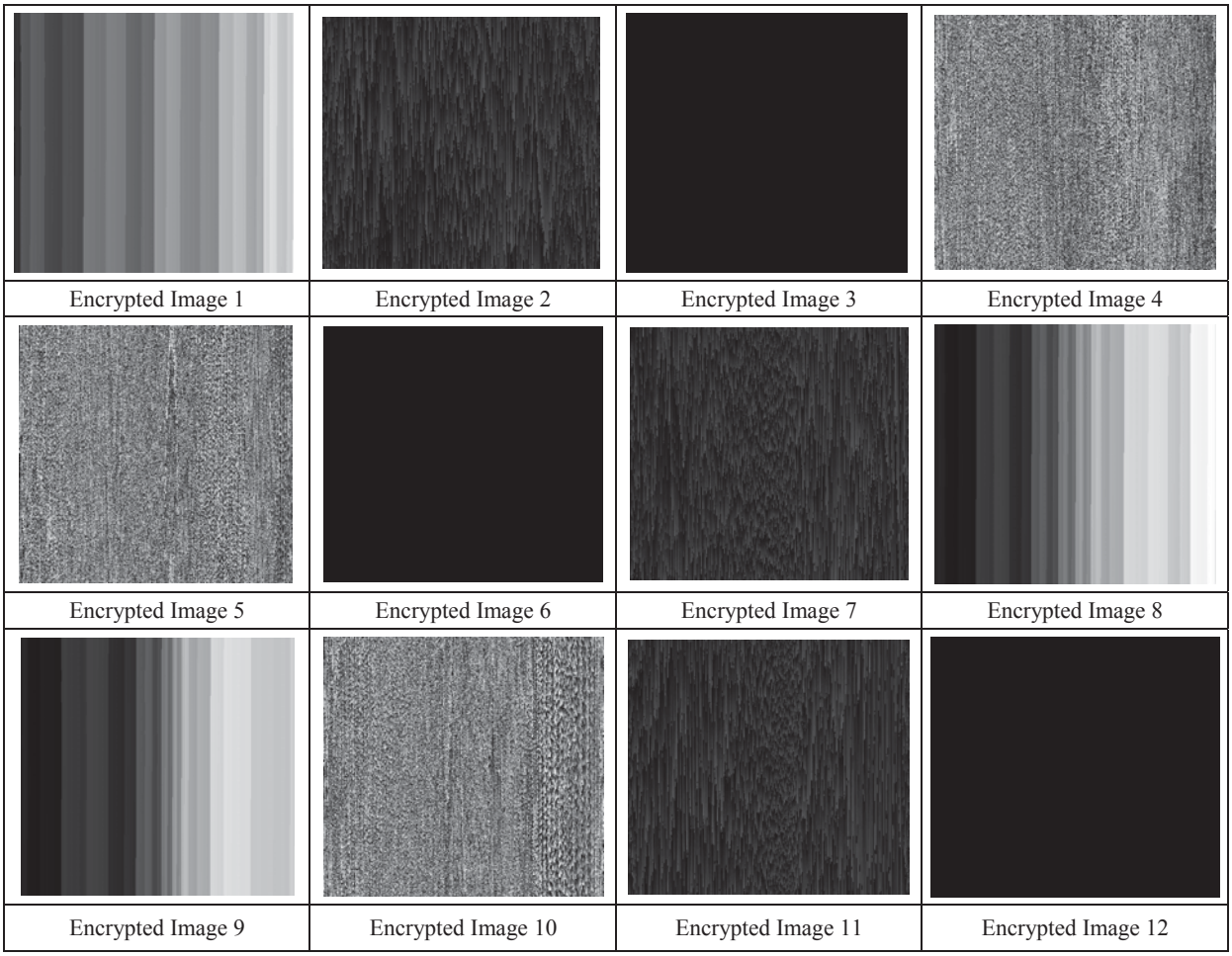


Fig. 5: encrypted image components

4. Experimental Results Analysis

A good encryption procedure should be robust against all kinds of cryptanalytic, statistical and brute-force attacks. Thus, the histogram of the encrypted images must be different from the original one to avoid statistical attacks, and the key space must be large enough to avoid brute force attacks. The M-PEST technique achieved the previous encryption requirements. Below performance analysis of the M-PEST technique shows that it is indeed robust against possible attacks.

4.1 Histogram Analysis

Visual attacks are the simplest and most important types of steganalysis. In a visual attack, the cipher-image is examined with the naked eye to identify any obvious inconsistencies. Figure 6 shows the secret image components, their corresponding cipher images and their histograms. It is clear that the histograms of the encrypted images are considerably different from the histogram of the original image. So the encrypted image does not provide any indication to utilize any statistical attack on the M-PEST, which makes statistical attacks difficult.

4.2 Key Space Analysis

The M-PEST technique has large key space that making brute force attack is infeasible. The M-PEST technique has different combinations of the secret keys. A cipher with such as a long key space is sufficient for practical use in secure cryptosystem and it's also threshold value.

Thus, the M-PEST technique makes larger statistical changes in the transmitted images and more secure. Experimental results demonstrate that proposed technique can defeat many existing steganalytic attacks. It is also have higher level of security against some existing attacks based on the multi-participant image index and encrypted keys and transposition and rotation.

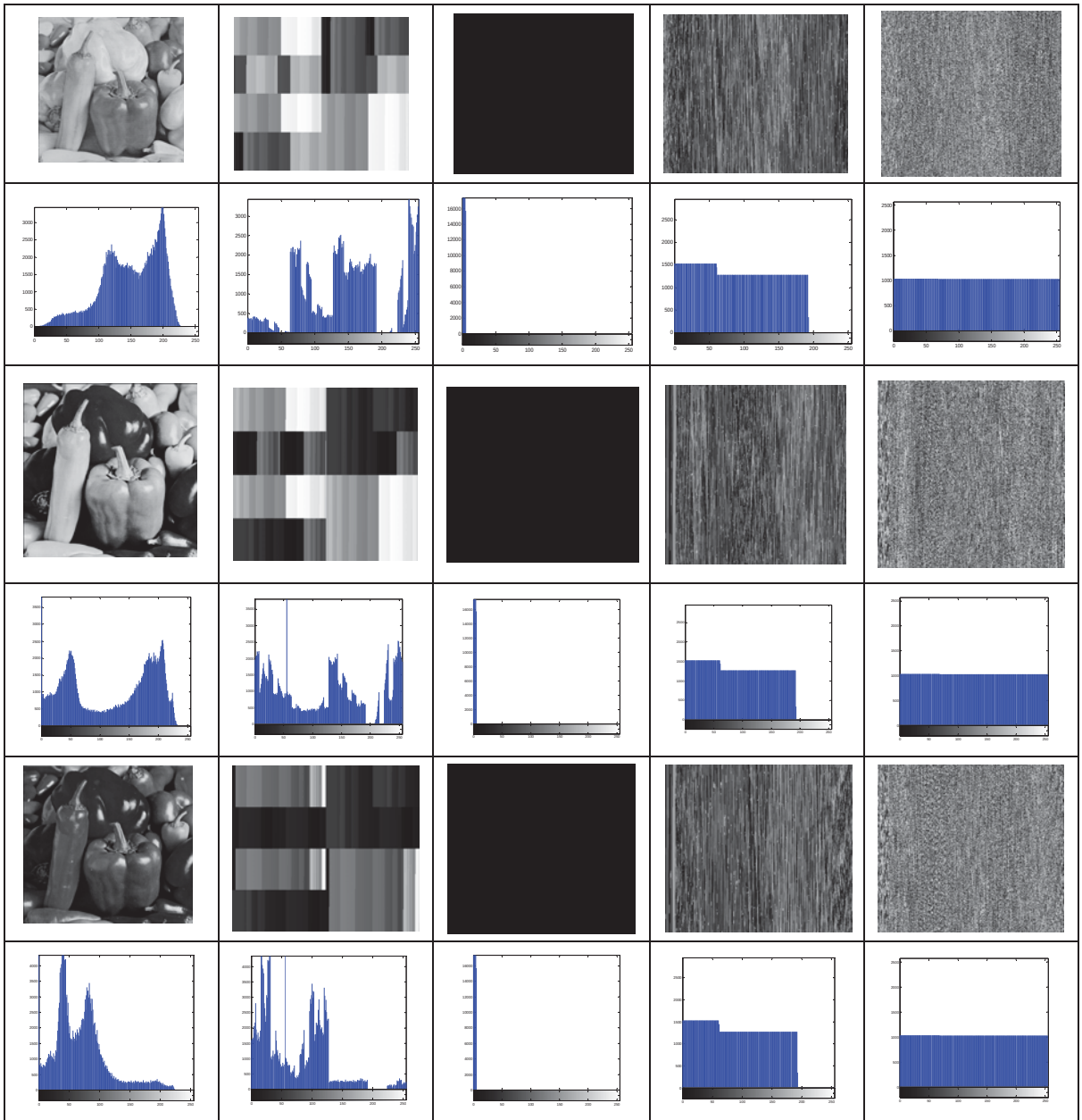


Fig. 6: Secret Image components, Encrypted Images and their histograms

5. Conclusions

In this paper, multi-participant encryption-steganography technique (M-PEST) is implemented for gray and color images. The M-PEST technique that is more secured encryption-steganography technique based on multi-participant images. This technique embeds the sensitive information into host image using bit exchange and divide index to multi-index matrices. It is also encrypting, rotating and transposing some blocks of the sorted secret image and its indices.

The results show the resistance of the M-PEST technique against different steganalytic attacks based on very large key space and different combinations. The changes of histograms of the secret image and its encrypted-images are

significant. The mean-squared error value between the original images and reconstructed images is zero.

Furthermore, they show it has multilayer protection stages and achieves confidentiality and gives more security, effectiveness and robustness to data and protects against detection.

In future work, we will be extended M-PEST technique for video and audio encryption. We try to use proposed technique to reduce the transmitted images size and enhance the encrypted-images quality.

REFERENCES

[1] William Stallings,"Cryptography and Network Security: Principles and Practice", Prentice Hall, 2011.

[2] Zaidoon Kh. AL-Ani, A.A.Zaidan, B.B.Zaidan and Hamdan.O.Alanazi, "Overview: Main Fundamentals for Steganography", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, pp. 158-165, MARCH 2010.

[3] Amitava Nag, Sushanta Biswas, Debasree Sarkar, and ParthaPratim Sarkar, "Semi Random Position Based Steganography for Resisting Statistical Steganalysis", International Journal of Network Security, Vol.17, No.1, PP.57-65, Jan. 2015

[4] M. Naor, A. Shamir, in: A. De Santis (Ed.), Visual Cryptography, Advances in Cryptology: Eurpocrypt'94, Lecture Notes in Computer Science, Vol. 950, Springer, Berlin, pp. 1–12, 1995.

[5] Anjali Varshney and Dinesh Goyal," Analysis and Design of Multi Share Secret Message Sharing using Visual Cryptography", International Journal of Computer Applications (0975 – 8887) Volume 102– No.13,pp. 19-23, September 2014

[6] Mr. ROHITH S, Mr. VINAY G, " A Novel Two Stage Binary Image Security System Using (2,2) Visual Cryptography Scheme", International Journal Of Computational Engineering Research,(IJCER), Vol. 2, Issue No.3, pp. 642-646 May-June 2012

[7] Chandramathi S., Ramesh Kumar R., Suresh R. and Harish S., "An overview of visual cryptography," International Journal of Computational Intelligence Techniques, ISSN: 0976–0466 & E-ISSN: 0976–047 vol. 1, Issue 1, pp.32-37, 2010.

[8] Omprasad Deshmukh and Shefali Sonavane," Multi-Share Crypt-Stego Authentication System", International Journal of Computer Science and Mobile Computing Vol.2 Issue. 2, pp. 80-90, February- 2013

[9] JagdeepVerma, Dr.VineetaKhemchandani, "A Visual Cryptographic Technique to Secure Image Shares," IJERA, vol. 2, Issue 1, pp.1121-1125, Jan-Feb. 2012.

[10] M. Agnihotra Sharma and M. ChinnaRao, "Visual Cryptography Authentication for Data Matrix Codes," International Journal of Computer Science and Telecommunications, vol 2, Issue 8, pp. 58-62 Nov. 2011.

[11] Mr. Parjanya C.A and Mr. Prasanna Kumar M , "Advance Secure Multi-Owner Data Sharing for Dynamic Groups in the Cloud", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 3, pp. 72-78, March 2014

[12] Shyamalendu Kandar and Bibhas Chandra Dhara, " k-n Secret Sharing Visual Cryptography Scheme on Color Image using Random Sequence", International Journal of Computer Applications (0975 – 8887) Volume 25– No.11,pp. 6-11, July 2011

[13] Jayanta Kumar Pal, J. K. Mandal and Kousik Dasgupta, "A (2, n) visual cryptographic technique for banking applications," International Journal of Network Security & Its Applications (IJNSA), vol.2, no.4, pp. 118-127 Oct. 2010.

[14] E. Verheul and H. V. Tilborg," Constructions And Properties Of K Out Of N Visual Secret Sharing Schemes."Designs, Codes and Cryptography, 11(2), pp. 179-196, 1997.

[15] Chang-Chou Lin and Wen-Hsiang Tsai, " Secret image sharing with steganography and authentication ", Journal of Systems and Software, Volume 73, Issue 3, pp. 405-414, November–December 2004.

Design Consideration for Client Synchronization Methods in an Identity Resolution Service

Fumiko Kobayashi, John R. Talburt
Department of Information Science
University of Arkansas at Little Rock
2801 South University Ave. EIT 550
Little Rock, AR, USA

Abstract- *Entity identity information management (EIIM) systems provide the information technology support for master data management (MDM) systems. One of the most important configurations of an EIIM system is identity resolution. In an identity resolution configuration, the EIIM system accepts entity identity information and returns the corresponding entity identifier. In the original EIIM model, identity resolution was only a batch operation. After the model was extended with an Identity Resolution Service (IRS)* **[1]** *to decouple identity resolution from batch EIIM processing, identity resolution moved into the interactive realm with additional functionality. This paper discusses the design consideration for one of these additional functions specifically centered on synchronization of entity identity information amongst client systems. This paper outlines the need and importance of client synchronization along with a design for obtaining client synchronization.*

Keywords- Identity Resolution Service; Entity Identity Information Management; Syncronization; Master Data Management; Entity Resolution

1 Introduction

Identity resolution (IR) is the process of determining if an entity reference refers to the same entity as one of the entity identity structures (EIS) under management in an entity identity information management (EIIM) system. IR is sometimes called "entity recognition" because the system is being asked if the input entity reference can be recognized as one of the entities already under management.

ER is the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different objects [2]. Real-world objects are identified by their attribute similarity and relationships with other entities. Some examples of attributes for person entities are First Name, Last Name, and Social Security Number (SSN). For place or location entities the attributes might be Latitude, Longitude, Description, or postal address. ER has also been studied under other names including but not limited to record linkage [3], deduplication [4], reference reconciliation [5], and object identification [6].

ER is a key task in data integration where different systems or data sources provide information for a common set of entities. ER has its roots in customer relationship management (CRM) where it is often referred to as customer data integration (CDI) [7]. The need for accurate and efficient ER is a necessity with the amount of data that is able to be collected and stored with current levels of technology. ER research is also driven by the need to share entity identity information across independently governed organizations in many areas such as education, healthcare, and national security.

Previous IQ research [8] [9] has extended ER into the larger context of entity identity information management (EIIM) that includes the creation and maintenance of persistent data structures to represent the identities of external entities [2]. The overall goal of EIIM is to allow the ER system to achieve entity identity integrity, a state in which two conditions hold [10].

1. Each identity structures corresponds to one, and only one, real-world entity
2. Distinct identity structures correspond to distinct real-world entities.

Entity identity integrity is another way of stating the Fundamental Law of Entity Resolution [2] which requires that two entity references should be linked if, and only if, they are equivalent where equivalence means both reference the same real-world entity.

In the current model of EIIM, the configurations to maintain the EIS operate primarily in an offline batch mode. In general the EIIM model is focused on the processes necessary to achieve and maintain entity identity integrity of the EIS under management in systems identity knowledgebase (IKB). EIIM provides the tools to support the complete life cycle of identity information.

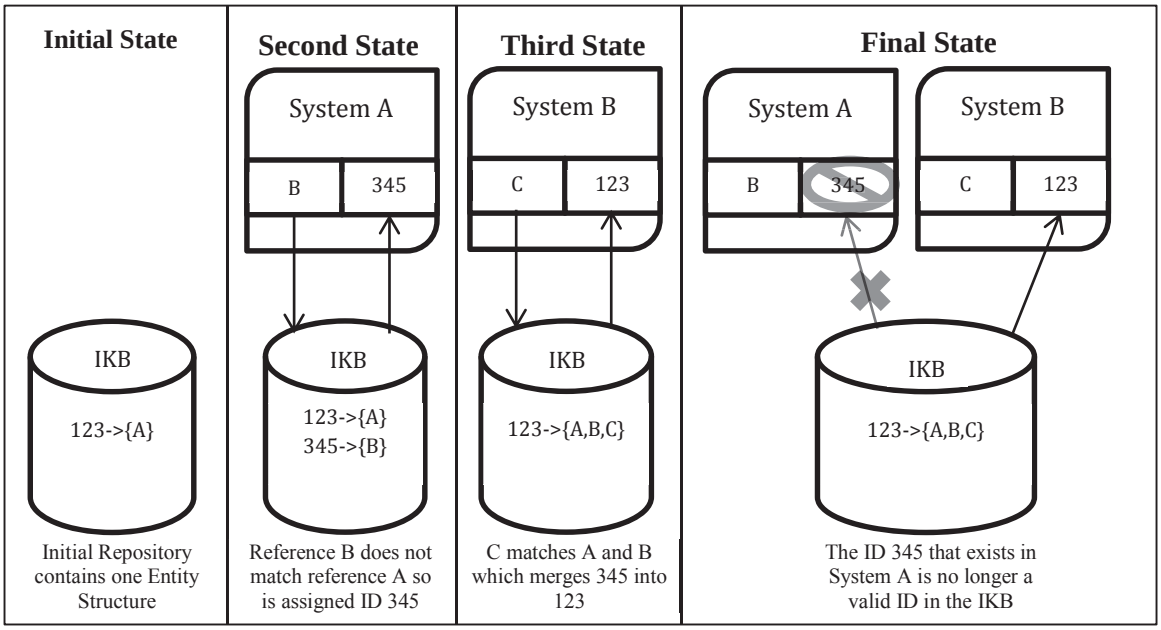


Figure 1: Identity State Issue in Current EIIM System

2 Problem Definition

The current EIIM model simply assumes that some external clients systems are providing entity identity information to the EIIM process, and in turn, the EIIM process is providing corresponding entity identifiers in the form of the link index table as described previously. However, no provision is made as to how these systems use and manage these entity identifiers. In particular, the issue of synchronization (consistency) is not addressed in the EIIM model. The problem is that when more than one system provides identity information to the EIIM system, identities in these systems can be inconsistently represented. For example, if System A provides an input record that is assigned entity identifier "345". Later System B provides new information causes the EIS assigned entity identifier "345" to be merged into the EIS with entity identifier "123". In this case, the identity "345" is no longer a valid entity identifier in the IKB; it has been superseded by entity identifier "123". This change will be reflected in System B that provided the information that caused the change, but System A is still using the entity identifier "345" for the same identity. This is illustrated in Figure 1.

There is no formal model for identity life cycle management which encompasses the client synchronization of EIIM systems to enforce persistent identification across various systems. Loss of client synchronization occurs when entity identifiers in the IKB change as the EIIM system performs updates to the centralized repository.

3 Identity Resolution Service (IRS)

An extension or new layer of functionality is required for the existing EIIM system. This new extended system is referred to as the Identity Resolution Service (IRS). Figure 2 shows that the additional layer added to an EIIM system creates the IRS that incorporates and maintains the functions of the EIIM system itself.

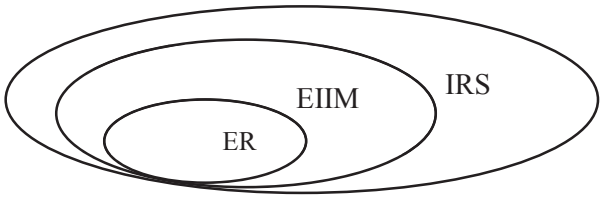


Figure 2: ER, EIIM, and IRS Relationship

The IRS can be invoked by applications from multiple remote client systems allowing them to retrieve identity interactively [1] [11]. By developing an IRS, clients are able to decouple and separate their identity management rules and processes from other business rules and processes.

The interactive IR system operates by accepting a reference(s) from the client system and then resolving the reference against the IKB. It accepts and returns one or more references depending on the mode of operation. If the IR system can identify matches, then the system returns the relevant identifiers to the client (along with other information depending on configuration). If the system is unable to find a positive match, then the system returns the most likely match along with a confidence rating. This confidence factor is calculated through the use of a probabilistic score algorithm [12] [13]. The confidence factor can help the client make a decision as to which possible match is the actual match (if any). If neither a match nor a set of possible matches can be identified, the system should inform the client or return an empty list to the client.

With the relocation of IR to the IRS layer, it is still tightly coupled with ER processing but allows for a different type of processing of the data. This IRS layer improves the IQ of the information by allowing it to be retrieved in a timely manner; by accessing the most up to date IKB so the data is not stale, and will even makes it easier for users to query the data which provides them a higher likelihood of identifying issues with the data and correcting it. Figure 3 shows the high-level components of an IRS.

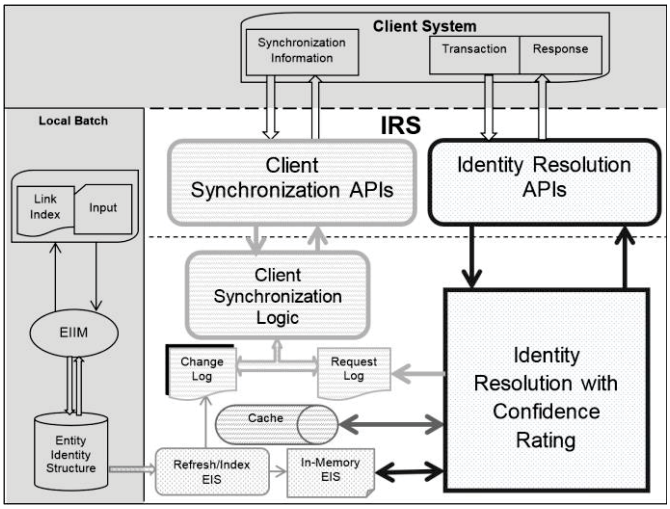


Figure 3: Identity Resolution Service

This paper discusses design for the client synchronization aspect of an Identity Resolution Service. An IRS is a web based service that allows multiple clients to access and use the identity information stored in the centralized IKB. It also allows a method to address stale information in the client systems without having to reprocess references against the IKB.

By introducing interactive identity resolution and a client synchronization framework, the IRS extends and enhances the EIIM system to fill the void that was left in regards to design for the Resolve and Retrieve Phase of the CSRUD (capture, store and share, resolve and retrieve, update, and dispose) MDM life cycle. For practical use, the Resolve and Retrieve Phase is the most important of all the of the CSRUD MDM life cycle phases. Resolving an entity reference to its correct entity (EIS) is the primary use case for MDM. It's this resolve and retrieval that provides actual value to the client systems.

A major issue for Resolve and Retrieve Phase is the client synchronization of entity identifiers in the centralized IRS system with entity identifiers residing in client systems. As entity identifiers change in the IRS IKB, the changes must be propagated to the clients' systems. This paper focuses on the need for and the design of the client synchronization functionality of an IRS.

4 Client Synchronization

Although there is no research directly addressing the identity synchronization in ER systems, the concept of synchronization has been studied for many other uses in the Information technology field. Data synchronization is the process of establishing consistency among data from a source to target data storage and vice versa. This consistency must be maintained between sources over time. Data synchronization is considered one of the key concepts behind most computing systems and networking protocols available. Data synchronization appears in many research fields such as distributed systems, mobile database [14], pervasive computing [15], mobile learning [16], data grids [17], and peer-to-peer content distribution [18].

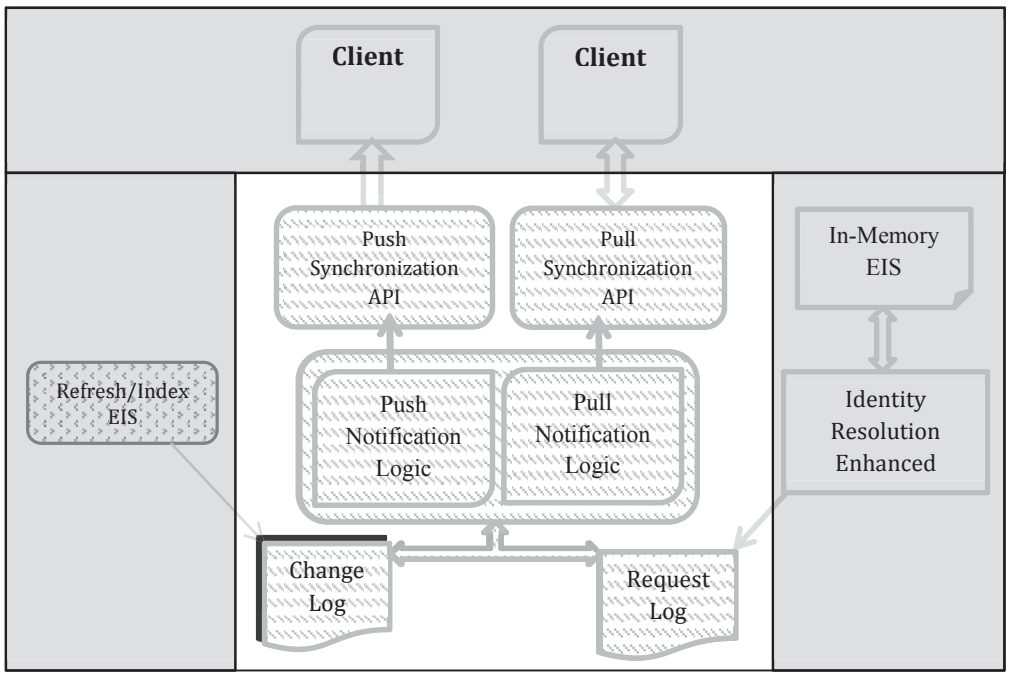


Figure 4: Push and Pull Synchronization

When performing client synchronization of the EIIM systems, the proper design of the client synchronization procedures is vital in insuring the correctness of the identities in all effected EIIM systems. Correctness of resolution is the main objective in ER systems and this importance must be carried through to synchronized EIIM systems.

Before design for the client synchronization methods can be done, the overall goal of an interactive system must be considered. One of the primary objectives in an interactive system is the timeliness of responses. By adding client synchronization methods to an interactive system, no impact to processing time should be experienced. To facilitate this, the use of logging is integrated into the client synchronization design. The final design encompasses two types of client synchronization, push and pull [19].

By including both models, it provides more robust methods of client synchronization to accommodate a wider variety of client systems.

4.1 Logging

The client synchronization design requires two types of logging, a change log and a request log.

The change log stores a list of all entity identifier values that have changed in the IKB along with the new entity identifiers value that they have changed to. These changes occur when a structure-to-structure assertion [20] run is performed by the EIIM system to bring multiple EIS into a single EIS. This also occurs when an input source references is found to be a "glue record" that brings multiple EIS together in the IKB. It is important to log these changes as it allows for accurate client synchronization of the IRS with remote client systems.

The request log stores a list of all the entity identifiers that have been sent to a client system along with ID associated to the client system.

4.2 Push Client Synchronization

In a push model, outlined in Figure 4, an update to the Change Log signifies a change to an EIS ID. When this occurs, the notification system queries the Request Log looking for any previous client systems that have requested or were returned the changed EIS ID. If a client system can be identified then the notification system will "Push" the updated information to the client system so that they can integrate it into their system and avoid having stale information. Figure 5 provides a more detailed example.

In this scenario the following occurred:

- An update occurred to the IKB that caused EIS E to merge with C
- The push notification logic noticed the new addition to the Change Log
- The IRS searched the Request Log for any request that had previously been provided ID E.
- Client system 1 was identified as having been provided this ID in the past
- The IRS system sends a push notice to Client system 1 alerting them to the update of ID

It is the responsibilities of the Client system to have a system in place that can receive and process these push notifications.

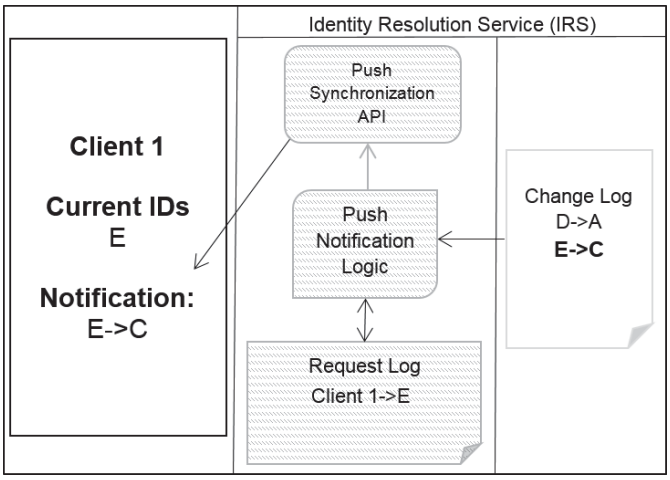


Figure 5: Push Client Synchronization Example

For the push model of client synchronization to function correctly, some sort of "registration" must occur between eh client system and the IRS. This is done via registration to the IRS and the assignment of an access key to the client. This works by allowing a client to register their system with the IRS and predefine the method for which they wish to receive notifications into their system. Once registration is complete the client will be assigned an access key which they must supply for every subsequent IR request into the system. The entity identifier that the client retrieved from the IRS is logged into the Request Log along with their access key. When an entity identifier is changed in the IKB, the Request log and access key are used to pick the correct connection and method for which to send the change notification.

For a fully optioned push client synchronization, two methods of notification are proposed. The first is the use of e-mail notifications. When the client system registers to receive e-mail notifications for changes, they must specify an e-mail address this is in turn associated with their access key in the IRS. When a change to an entity identifier occurs, an e-mail containing the following body is sent:

{OldID=XXXXXXXXXX;NewID=YYYYYYYYY}

The client can parse the e-mails and update their local IKB.

The second method of push client synchronization is done via an API call. The client system must have an API for which

the IRS can make a POST call. The body of the call will include an XML structure that conveys the change. This will look like this:

```
<Change_Notification>
    <OldID>XXXXXXX</OldID>
    <NewID>YYYYYYYY</NewID>
</Change_Notification>
```

The client system must be able to receive and process the xml notification posted to their API. The client must provide the web address of their API when registering for an access key in the IRS.

4.3 Pull Client Synchronization

In a pull model, depicted in Figure 4, the Client system submits an EIS IDs that was previously provided to them by the IRS service back to the system. The notification system will query the Change Log and identify any changes that occurred for the provided EIS ID. If a change is found, this information is provided to the Client system. By having the Logs in conjunction with the notification framework, it is possible to provide Client systems with relevant information beyond what is possible using the IKB alone. Figure 6 provides an example of how this notification configuration functions in the IRS.

In this scenario the following occurred:

- The Client system submitted a request to check if ID for EIS E has changed
- The pull notification logic searches the Change Log
- It finds that EIS E has changed and merged with EIS C
- The IRS returns this change to the client system.

This may seem similar to how the ID Search feature works but with one significant different. This request requires much less processing as the in-memory IKB is never touched to find the answer. Only a search against the smaller indexed change log is required. If the client system later decides they would like the full content of the new EIS, they will have to submit a new ID Search request.

In a pull synchronization, no access key is required. This is a process that is strictly initiated from an API call from the client system. The client sends an entity identifier they have in their IKB and the system sends them back the following if a change to the identifier is located:

{OldID=XXXXXXXXXX;NewID=YYYYYYYYY}

If no change is identified, the system will return a null set like:

{NULL}

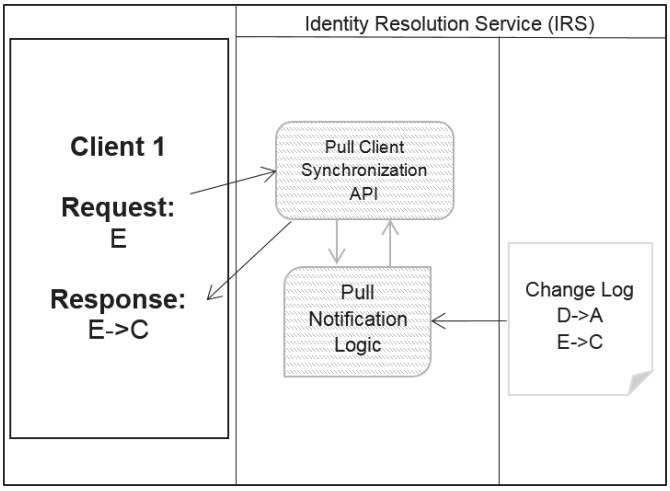


Figure 6: Pull Notification Example

5 Summery

The design of a client synchronization and notification framework extends the entity identity life cycle management model to allow for client synchronization of entity identifiers across client systems.

The separation of IR from the EIIM system to improve access to the information stored in the identities creates the problem of lost client synchronization. A logical extension to the EIIM model is a client synchronization maintenance task to allow for the information received by clients to remain current if the data in the Knowledgebase is updated. This client synchronization is designed for inclusion into the IRS.

The client synchronization services, used by Client 1 in Figure 5 and Figure 6, provide a mechanism for the clients to keep the freshest identifiers for their identities and stay in sync with the centralized IKB. By including both models for client synchronization, push and pull, in the system.

In a more general sense a pull model becomes the users' responsibility to inquire into the management system to determine if change has occurred or not to their identities in their local IKB. A pull notification would provide applications developers the ability to send an identifier that is in use in their system and check if it is still in sync with the centralized repository without having to perform resolution on the reference again.

One possible model for the client synchronization service built as a push notification would be a feature of the Identity Service which pushes a combination of old identifier and new identifier to clients that have made requests to the system in the past. It is up to the client to process the update notification when it is received and update their information. The push model is a more complex model which places the responsibility on the management system to remember which clients, and which client references have been processed and to log changes and push the changes back to the client system

No matter which model is selected for use, the benefits of a client synchronization framework are clear in that it solve the issue of mismatched identity state that exists in the current EIIM model. It also adds value to the IRS and improves the accuracy of the data stored in the client IKB since there is a method to address and update stale data.

6 Future Work

Future work for a client synchronization frame work will consist of extensive testing of the frameworks proposed in this paper. This experimentation will be aimed at providing validation that the proposed models are viable for a production system. Issues surrounding the concepts of commit and rollback of entity identifier synchronization must also be addressed.

Acknowledgment

The research described in this paper has been supported in part by funding from the Arkansas Department of Education.

References

[1] Fumiko Kobayashi and John R Talburt, "Decoupling Identity Resolution from the Maintenance of Identity Information," in *International Conference on Information Technology (ITNG)*, Las Vegas, NV, 2014.

[2] John R. Talburt, *Entity Resolution and Information Quality*. Burlington, MA: Morgan Kaufmann, 2011.

[3] Howard B Newcombe, James M Kennedy, S J Axford, and A P James, "Automatic Linkage of Vital Records," *Science*, vol. 130, pp. 954--959, October 1959.

[4] Sunita Sarawagi and Anuradha Bhamidipaty, "Interactive deduplication using active learning," in *KDD '02*, Edmonton, Alberta, Canada, 2002, pp. 269--278.

[5] Xin Dong, Alon Halevy, and Jayant Madhavan, "Reference reconciliation in complex information spaces," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, 2005, pp. 85--96.

[6] Sheila Tejada, Craig A Knoblock, and Steven Minton, "Learning object identification rules for information integration," *Inf. Syst.*, vol. 26, pp. 607--633, dec 2001.

[7] J Dyché, E Levy, D Peppers, and M Rogers, *Customer Data Integration: Reaching a Single Version of the Truth*.: Wiley, 2006.

[8] Richard Y Wang and Diane M Strong, "Beyond accuracy: what data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, pp. 5--33, mar 1996.

[9] Yinle Zhou and John Talburt, "Entity identity information management (EIIM)," in *2011 International Conference on Information Quality (IQIC11)*, Australia, 2011, pp. 327-341.

[10] Yinle Zhou and John Talburt, "The Role of Asserted Resolution in Entity Identity Management," in *2011 International Conference on Information and Knowledge Engineering (IKE'11)*, Las Vegas, Nevada, 2011, pp. 291-296.

[11] Fumiko Kobayashi, Eric Nelson, and John D. Talburt, "DESIGN CONSIDERATION FOR IDENTITY RESOLUTION IN BATCH AND INTERACTIVE ARCHITECTURES," in *Proceedings of the 16th International Conference on Infromation Quality (ICIQ-11)*, Adelaide, Australia, 2011, pp. 287-300.

[12] Fumiko Kobayashi and John R Talburt, "Probabilistic Scoring Methods to Assist Entity Resolution Systems Using Boolean Rules," in *Proceedings of the 2013 International Conference on Information and Knowledge Engineering (IKE'13)*, Las Vegas, NV, 2013.

[13] Fumiko Kobayashi and John R Talburt, "Deciding Confidence for Identity Resolution in Closed and Open Universes of Entity Identity Inforamtion," in *The Fifth International Conference on Business Intelligence and Technology (BUSTECH 2015)*, Nice, France, 2015.

[14] Yang Li, Xuejie Zhang, and Yun Gao, "Object-Oriented Data Synchronization for Mobile Database over Mobile Ad-hoc Networks," , vol. 2, 2008, pp. 133-138.

[15] Yun-Wu Huang and P S Yu, "Lightweight version vectors for pervasive computing devices," , 2000, pp. 43-48.

[16] V Tam and Barbara Yin, "Investigating data synchronization in a mobile learning network with handheld devices," , 2003, pp. 296-300.

[17] Srikumar Venugopal, Rajkumar Buyya, and Kotagiri Ramamohanarao, "A taxonomy of Data Grids for distributed data sharing, management, and processing," *ACM Comput. Surv.*, vol. 38, jun 2006.

[18] Stephanos Androutsellis-Theotokis and Diomidis Spinellis, "A survey of peer-to-peer content distribution technologies," *ACM Comput. Surv.*, vol. 36, pp. 335--371, dec 2004.

[19] Jean-Philippe Martin-Flatin, "Push vs. Pull in Web-Based Network Management," in *IM'99*, Boston, MA, 1999.

[20] Yinle Zhou and John Talburt, "The Role of Asserted Resolution in Entity Identity Management," in *The 2011 International Conference on Information and Knowledge Engineering (IKE'11)*, Las Vegas, Nevada, 2011.

Applying Phonetic Hash Functions to Improve Record Linking in Student Enrollment Data

(Research in progress)

A. Pei Wang[1], B. Daniel Pullen[2], C. John Talburt[2] and D. Ningning Wu[2]
[1]Department Information Science Department
University of Arkansas at Little Rock
Little Rock, AR, USA

Abstract: *Entity resolution and record linking processes are often required to process input records of poor data quality. However, the matching errors caused by poor quality data can often be overcome by categorizing the quality problems, then applying a cyclic process that continuously refines the match rules to overcome these problems. This paper presents an extension to a previous case study of this process for student enrollment data and describes how the unique data quality issues that were identified throughout this cyclic process and how different phonetic hashing functions were used to overcome these issues.*

Key Word: Entity Resolution, Record Linkage, Phonetic Hash Code, Data Quality (DQ), Boolean matching rules

1. Introduction

Previous work in this area has been published utilizing similarity functions such as Levenshtein Edit Distance and Q-gram Tetrahedral Ratio [11]. This research takes a different approach by applying phonetic hash code functions to mitigate quality issues presented in student enrollment data. This approach can help to overcome variations that stem from phonetic to text conversion performed by humans as well as overcome common typographical variations.

2. Background

Entity Resolution (ER) is the process of determining whether two references to real world objects in an information system are referring to the same object or to different objects [1]. The references are made up of attributes and the values of the attributes describe the real world entity to which they refer. The ER processes discussed in this paper use Boolean match rules to make their decisions. Boolean match rules do not produce a score or weight when comparing a pair of references, only a True/False decision. If two references satisfy a Boolean match rule, i.e. the rule is "true", the references are linked together. After the application of transitive closure, all of the references that can be linked together form an entity identity structure (EIS) [9].

3. Boolean Match Rules and ER Outcomes

Boolean match rules are used to determine the outcome as "link" pairs or "non-link" pairs. The basic unit of a Boolean rule is a term. A term is the comparison between the values of an attribute in the pair of records. The term is considered to be "TRUE" if the degree of similarity required by the comparison is met. The rule itself is made up of a series of terms connected by "AND" logic, i.e. every term must be true in order for the rule to be true. Finally, the ER process may use several Boolean rules that are connected by "OR" logic, i.e. the pair of references should be linked if at least one of the Boolean rules is true [3].

In evaluating the outcome of an ER process, the results of the matches between all pairs of references can be placed into four, mutually exclusive categories: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). TPs are correctly labeled "link" pairs. TNs are correctly labeled "non-link" pairs. Contrasting these correct results are two types of incorrect linking results. FPs are pairs of records that have been identified as matches or "link" pairs by the ER process but actually refer to two

different real world entities. FNs are pairs of records that have been identified as non-matches or "non-link" pairs by an ER process but actually refer to the same real world entity [8]. The goal of an ER process is to produce the lowest number of FPs and FNs.

4. The OYSTER ER System

The ER processes in this paper were performed with OYSTER (Open sYStem for Entity Resolution). OYSTER is an open source ER system developed by the Center for Advance Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at Little Rock (UALR). OYSTER was specifically designed to support entity identity information management (EIIM) [9]. Although OYSTER can be run in several different configurations to support the various phases of the entity identity information life cycle, only the identity capture configuration was used for the results given in this paper [10].

ER Impact of Data Quality Issues

The data set used throughout this testing is a collection of student enrollment data spanning two academic years. The total records and Clusters are listed in Table1.

TABLE I. DATA SETS

	Set A	Set B
Total Cluster	526,362	426,934
Total Records	3,234,292	3,255,513

Only the student identity information was used. Any results discussed in this paper have been made anonymous to allow the sharing and description of the unique cases identified. In the data available, a few strong identifying attributes are of particular interest. These are first name, middle name, last name, date of birth, and student identifier. Some of the data quality (DQ) issues identified with these attributes and their rates are summarized below in Tables 2 and Tables 3.

TABLE II. DATA QUALITY ISSUES IN DATA SET A

Data Quality Issue	Data Set A	%
Number in First Name	121	0.003741
Number in Middle Name	165	0.005102
Number in Last Name	35	0.001082
Virgule in First Name	24	0.000742
Asterisk in First Name	93	0.002875
Total Problems	438	0.013542
Total Records	3,234,292	

TABLE III. DATA QUALITY ISSUES IN DATA SET B

Data Quality Issue	Data Set B	%
Number in First Name	135	0.004147
Number in Middle Name	136	0.004178
Number in Last Name	31	0.000952
Virgule in First Name	27	0.000829
Asterisk in First Name	66	0.002027
Total Problems	395	0.012133
Total Records	3,255,513	

These tables point out some of the obvious and easily quantifiable data quality issues present in these two data sets. There are several other data quality issues that occur over these attributes. The student name fields have some particularly interesting and challenging problems. The fields occur frequently enough throughout the data set to increase the amount of errors made by the ER process.

The first name field has many records where the field is treated not only as the student's first name but also nickname. This creates examples that look like "Joseph (Joey)" or "Joseph Joey." In other cases, Many Hispanic students have a hyphenated name where one comes from the father and the other comes from the mother. Upon data entry, sometimes the first of the two names is placed in the middle name field. This has a detrimental impact on matching using the middle and last name fields. In addition, to the presence of numbers or special characters in all three of the name fields can cause problems.

Some problems affect multiple attributes. Some of these unique cases can be summarized briefly. In some cases one attribute is placed in the incorrect

field. Cases involving the phone number, student identifier, and address field have been identified where these values are actually in one of the student name fields. The data also shows a trend in naming twins. Often parents will name the twins with very similar names such as "Terrell" and "Jerrell." Occasionally, this is extended to a similarity in the middle names as well. With the date of birth and last name fields already identical, differentiating twins in the match rules is problematic. In some cases, mixing this with erroneous or sequential student identifiers can create FP outcomes.

5. Methodology

How can managers of entity data overcome data quality problems when performing ER? To overcome data quality issues some appropriate similarity functions and comparator functions can make a notable improvement.

IBM Alpha Code - IBM Alpha Code is a name encoding algorithm. The coding rules produce a 14 digit phonetic key of the name according to some rules [11]. Based on these phonetic keys, the name which has different spelling but same pronunciation can be matched to each other. For example: value 1 = "Rodgers" and value 2 = "Rogers".

The New York State Identification and Intelligence System (NYSIIS) - It is a phonetic algorithm. Much like the previous algorithm, a name with different spelling can produce a match by using this function. For example: value 1 = "Carry" and value 2 = "Carrie".

Soundex – Soundex can be used to find the values which have similar pronunciation but difference spelling. This function can be used to fix misspelled and even transposed characters. For example value 1 = "Damieva" and value 2 = "Dameiva." These two values will produce the same Soundex hash value, creating a match.

Scan – In order to overcome the special characters in names, the similarity function scan can be used. It is often performed in preprocessing before the ER is completed and has the capability to filter all the special characters and only include letters or alphanumerical characters. For example, value 1 = "JAMES\\" and value 2 = "JAMES". Also, scan can reorder strings or even read them from right to left as opposed to left to right and perform transformations regarding the casing of alphabetical characters. This comparator can force all characters to be lower case, upper case, or the original case present in the string. For example "Eric" can be generated as "ERIC" after using scan.

Sometimes, these similarity functions will create FPs and FNs. For example, suppose two different rules are used to produce two different ER results from the same data set. The first rule we use is student first name, student last name and date of birth with an exact match for each of them. The second rule is student first name Soundex, last name and date of birth with an exact match. After performing a split comparison to compare the two results as in previous research [7], the FPs and FNs created by the second rule can be identified and their rates can be calculated. The calculation for the approximate FP percentage rate is shown in equation (1). The results are shown in table 3. Since these FPs were identified using split analysis, these are considered to be worst case FP rates. Split analysis is a methodology used to analyze splits in the clusters between two different link identifiers. How this process works has been discussed in detail in recent research [7].

$$FP\% = \frac{FP\ Count}{Linked\ Count}(100) \qquad (1)$$

The FP rate indicates one side of how well the rules are performing. For this reason, the user should attempt to reduce FP and FN rates as low as possible when creating and testing rules. These results focus on the FN rates in particular.

This research focuses on three similarity functions. These are Soundex, NYSIIS and IBM Alpha Code. These three similarity functions can be used in indexing, which can help the process to speed up, especially for the large data sets. After testing these three functions in the same student enrollment data, the percentage of TP and FP are shown in the table below (Table 4):

TABLE IV. THE PERCENTAGES OF TP AND FP

	TP	FP	Not Sure
Soundex	34.6%	62.5%	2.9%
NYSIIS	36.5%	56.9%	6.7%
IBMAlpha	27.4%	68.0%	4.6%

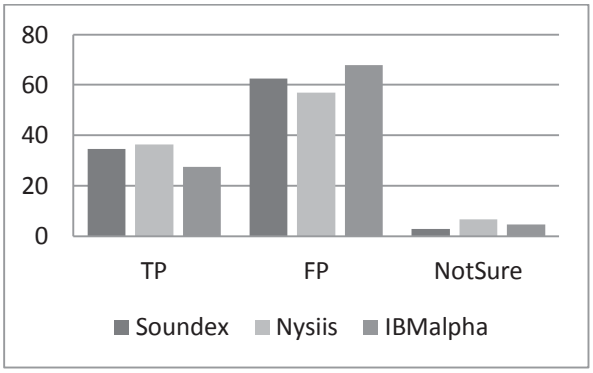


Fig. 1. Bar Graph of TP and FP Percentages

Comparing the results of these three similarity functions to benchmark, clearly the one that has the best performance is NYSIIS, which has the highest percentage of TP and lowest percentage of FP.

6. Conclusions

Data quality problems often present a formidable obstacle to obtaining an accurate and effective ER result. The approaches to overcome data quality issues in the student enrollment data during ER described in this paper have been successfully implemented in OYSTER. The success of any ER process is often directly related to the time spent profiling the data and identifying these types of data quality problems. Effectively identifying and categorizing these types of problems directly affect the quality of the ER results at the end of such processes.

The approaches above include the similarity functions such as Soundex and IBMAlphaCode that can overcome some issues such as both nickname and given name contained together in one field, transposed characters, and other typographical or spelling errors. Additionally, other similarity functions such as Scan can overcome the issues such as special characters, numbers, and misspellings. While these approaches contribute greatly to improving the ER results, there is a limit to which of these approaches can aid in reducing the FP rate.

The hash code functions tested in this paper cannot overcome all of the issues listed earlier in this paper. For example, they cannot directly overcome variations produced by the inclusion of nickname in some references but a given name in other references. However, the application of these hash code functions along with similarity functions such as q-gram tetrahedral ratio, Levenshtein edit distance, and nickname could further mitigate the issues encountered in this particular student enrollment data.

7. Acknowledgment

The research described in this paper has been supported in part through grants from the Arkansas Department of Education and Black Oak Analytics.

8. Reference

[1] Talburt, John R. Entity Resolution and Information Quality. San Francisco, CA: Morgan Kaufmann/Elsevier, 2011.

[2] Melody Penning and John Talburt. "Information Quality Assessment and Improvement of Student Information in the University Environment". Information and Knowledge Engineering, 2012.

[3] Yinle Zhou, John Talburt, Fumiko Kobayashi and Eric D.Nelson. "Implementing Boolean Matching Rules in an Entity Resolution System using XML Scripts". Information and Knowledge Engineering, 2012.

[4] Holland, G. & Talburt, J. (2010) q-Gram Tetrahedral Ratio (aTR) for approximate pattern matching. 2010 Conference on Applied Research in Information Technology, University of Central Arkansas, Conway, AR.

[5] IvenFellegi and Alan Sunter. “A Theory for Record Linkage”; Journal of the American Statistical Association, Vol. 64 No. 328, 1183-1210, 1969

[6] Steven Whang and Hector Garcia-Molina. “Entity Resolution with Evolving Rules”; Proceedings of the VLDB Endowment, Vol. 3 Issue 1-2, 1326-1337, September 2010

[7] Huzaifa Syed, Fan Lui, Daniel Pullen, Ningning Wu, John Talburt. "Developing and Refining Matching Rules for Entity Resolution"; Information and Knowledge Engineering, 2012

[8] Christen, Peter. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer, 2012.

[9] Zhou, Y. and Talburt, J. (2011). Entity Identity Information Management (EIIM). International Conference on Information Quality (ICIQ-11), Adelaide, Australia, November 18-20, 2011, pp. 327-341

[10] Zhou, Y. and Talburt, J. (2011). The Role of Asserted Resolution in Entity Identity Management. The 2011 International Conference on Information and Knowledge Engineering (IKE'11), Las Vegas, Nevada, July 18-20, 2011, pp.291-296

[11] Wang, Pei, Pullen, Daniel, Wu, Ningning, and Talburt, John. (2013) Mitigating Data Quality Impairment on Entity Resolution Errors in Student Enrollment Data; Information and Knowledge Engineering Conference, 2013.

SESSION

LATE BREAKING PAPERS: SMART CITY ECOSYSTEMS, UNL, AND ENVIRONMENTAL ISSUES

Chair(s)

TBA

Lifecycle Based Modeling of Smart City Ecosystem

Ahmed Hefnawy[1,3], **Abdelaziz Bouras**[2,3], **Chantal Cherifi**[1]
1 DISP Lab, Lyon 2 University, Lyon, France
2 DCSE, College of Engineering, Qatar University, Qatar
3 Ministry of Information and Communications Technology (ictQATAR), Qatar

Abstract - *Smart city services have an inevitable role in addressing the complexity of modern city operation. Smart transport, smart parking, smart energy, smart water and many others are examples of vertical smart city systems that are mainly concerned with its particular domain. Realizing the full promise of smart city will require interoperability among those systems and data fusion between heterogeneous components from different domains. In this regard, many standardization organizations have been working on modeling smart city and similar or related systems and concepts, such as Internet of Things (IoT) and Cyber Physical Systems (CPS), to ensure common technical grounding and architectural principles. Though, there is still a need to address the higher-level requirements of smart city as a complete ecosystem. To this end, this paper discusses different Smart City solutions and highlights lifecycle based modeling to better integrate people, processes, and systems; and assure information consistency, traceability, and long-term archiving.*

Keywords: Smart City; IoT; CPS; Data Fusion; Lifecycle Management.

1. Introduction

The world is witnessing continuous global tendency towards urbanization. The world's population residing in urban areas has increased from 30 percent in 1950 to 54 percent in 2014 and forecasted to reach up to 66 percent by 2050. In addition, by 2030, the world is expected to have 41 mega-cities with more than 10 million inhabitants [1]. On one hand, high concentration of population empowers cities and fuels economic growth. On the other hand, significant challenges of sustainability and complex city operation are likely to accompany advantages of urbanization. The increasing complexity of traffic congestions, waste management, human health concerns, environmental pollution, scarcity of resources and inefficient allocation makes ordinary service provisioning less effective compared with innovative smart city services [2].

The International Telecommunication Union (ITU) defines a smart sustainable city as *"an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services…"* [3]. The British Standards Institution (BSI) was even more specific when described this innovative smart city as *"an effective integration of physical, digital and human systems in the built environment to deliver a sustainable, prosperous and inclusive future for its citizens"* [4]. The integration of physical and digital/ cyber systems, in co-engineered interacting networks, is widely known as "Cyber -Physical Systems" (CPS) [5]; or similarly, as "Internet of Things" (IoT) which is defined as "The global network connecting any smart object" [6]. The global connectivity feature of CPS and IoT fuels smart city with real-time data streams about certain characteristics of the real world [5]; and hence, smart city services empower city operators with real-time decision-making enabled by real-time data streams from heterogeneous objects. Smart transport, smart parking, smart energy, smart water are just few examples of smart city systems. Bearing in mind that the mentioned systems address sector-specific challenges; the resulting smart city applications appear as vertical silos, locked to specific domains, with less consideration to collaboration between those vertical silos.

In this regard, many standard organizations have been working on modeling smart city, IoT and CPS, to ensure common technical grounding and architectural principles. Though, there is still a need to address the higher-level requirements of smart city as a complete ecosystem. In fact, the smart city ecosystem is wider than only technical systems. The ecosystem equally includes human, whether users, policy makers, regulators, vendors, etc. The ecosystem has also business models and processes; and subject to applicable laws, policies and regulations. Finally yet importantly, the smart city ecosystem is more about the entire quality of life and living standards rather than isolated experiences in one or more sectors. Therefore, the objective of this paper is to consider high-level requirements of the smart city ecosystem in order to ensure horizontal flow of valuable information between multiple stakeholders, across different domains. S. Kubler, K. Främling, et al. argue that this concept is closely linked to Lifecycle concepts, which is commonly understood as a strategic approach that incorporates the management of data, versions, variants and business processes associated with heterogeneous, uniquely identifiable and connected objects [7][8].

This paper proposes lifecycle based modeling of the entire smart city ecosystem to ensure systematic involvement and seamless flow of information between different stakeholders of the smart city ecosystem. The remaining of this paper is structured as follows: Section 2 describes the Smart City Framework (SCF), and other relevant concepts/ views of CPS and IoT models. Section 3 explains the proposed high-level approach of lifecycle based modeling of smart city ecosystem. Section 4 discusses the proposed approach and the applicable lifecycle management systems. Section 5 sheds light on the conclusion of this paper and the proposed future work.

2. Smart cities reference models

Many standardization and research institutes are currently working on standardizing and modeling Cyber Physical Systems (CPS), Internet of Things (IoT) and Smart Cities. NIST is currently leading the work on CPS through the CPS Public Working Group (CPS PWG). From 2010 to 2013, the European Lighthouse Integrated Project "Internet of Things – Architecture" (IoT-A) developed an architectural reference model for the IoT, referred to as the IoT Architectural Reference Model (IoT-A-ARM). In 2014, the IEEE established P2413 working group with the scope to define an architectural framework for the Internet of Things (IoT). From 2013 to 2016, the CityPulse project has been working on Smart City Framework (SCF) to serve as a Reference Architecture Model [9]. The undergoing work in modeling of smart city, IoT and CPS, is very comprehensive and massive. For the purposes of this paper, this section focuses on SCF as the most currently available prominent reference model of smart city. This section also presents the IoT functional model, since SCF uses IoT sensors and actuators as one type of information sources and sinks respectively. Finally, this section presents the concepts of Lifecycle Management in the context of CPS.

2.1. Smart City Framework

The purpose of the Smart City Framework (SCF) is to set the main concepts, common language and the boundaries to be used by smart city stakeholders, partners and interested parties when engaged in technical discussions about smart city services [9]. There are three main groups of SCF stakeholders: City Stakeholders (IT service providers, City departments and City decision makers); Third Party Providers (e.g. App developers); and Citizens. The high-level view of SCF, illustrated in Figure 1, has different interfaces (I/F) towards the applications and towards the information sources/ sinks. Information Sources include: Internet of Things (IoT) sensors deployed in a city environment; city information sources e.g. Open Data portals, city Geographical Information System (GIS) data etc.; and, user generated information through social media e.g. microblogs such as tweets that have been proven feasible for city related event extraction. Information Sinks include: IoT Actuators, City Datastores and social media channels through which cities could potentially push information to their citizens. The SCF consists of number of Functional Groups (FGs). The Large-Scale Data Analysis FG addresses issues related to integration of a large scale of heterogeneous sources producing real-time streams and their semantic enrichment. The Reasoning and Decision Support FG tackles issues related to the ability of the SCF to adapt to alterations based on real-time information streams. It is mainly responsible for monitoring the semantically enriched streams and adapting the collection of stream information from one side and providing an API towards the Smart City Applications from another side. The Large Scale Analysis and Reasoning and Decision Support functionalities are supported by prior knowledge in the form of the Knowledge Base FG and Reliability and Quality of Information control mechanisms by the Reliable Information Processing FG.

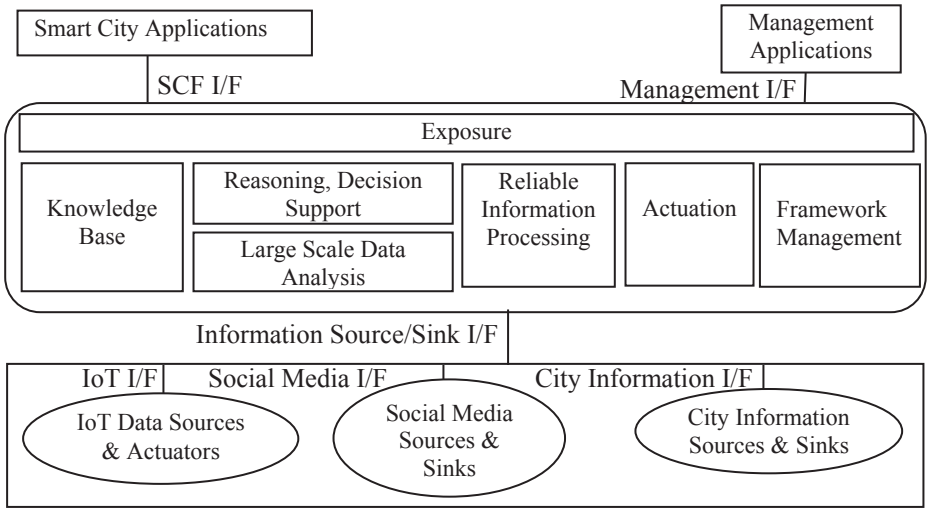


Figure 1: High-level view of Smart City Framework [9]

The Actuation FG covers any functionality that allows the SCF to push control commands or information to the IoT actuators, social media sinks and city information sinks. The Framework Management FG includes functionalities for the management of the SCF itself such as fault, configuration, security management etc. The Exposure FG covers the mediation of access with management and smart city applications.

2.2. Internet of Things approach

The IoT Functional Model, as proposed by the IoT-A project [6], illustrated in Figure 2, contains seven longitudinal Function Groups (FGs) (light blue) complemented by two transversal FGs (Management and Security, dark blue). The IoT Process Management FG relates to the conceptual integration of (business) process management systems with the IoT-A-ARM. The Service Organization FG is responsible for composing and orchestrating services of different levels of abstraction. It effectively links service requests from high level FGs such as the IoT Process Management FG, or even external applications, to basic services that expose resources and enables the association of entities with these services by utilizing the Virtual Entity FG. The Virtual Entity and IoT Service FGs include functions that relate to interactions on the Virtual Entity and IoT Service abstraction levels, respectively. The Virtual Entity FG contains functions for interacting with the IoT System on the basis of Virtual Entities, as well as functionalities for discovering and looking up services that can provide information about Virtual Entities, or which allow the interaction with Virtual Entities. Furthermore, it contains all the functionality needed for managing associations, as well as dynamically finding new associations and monitoring their validity. The IoT Service FG contains IoT Services as well as functionalities for discovery, look-up, and name resolution of IoT Services. The Communication FG provides a simple interface for instantiating and for managing high-level information flow. The Management FG combines all functions that are needed to govern an IoT system. The Security FG is responsible for ensuring the security and privacy of IoT-A-compliant systems.

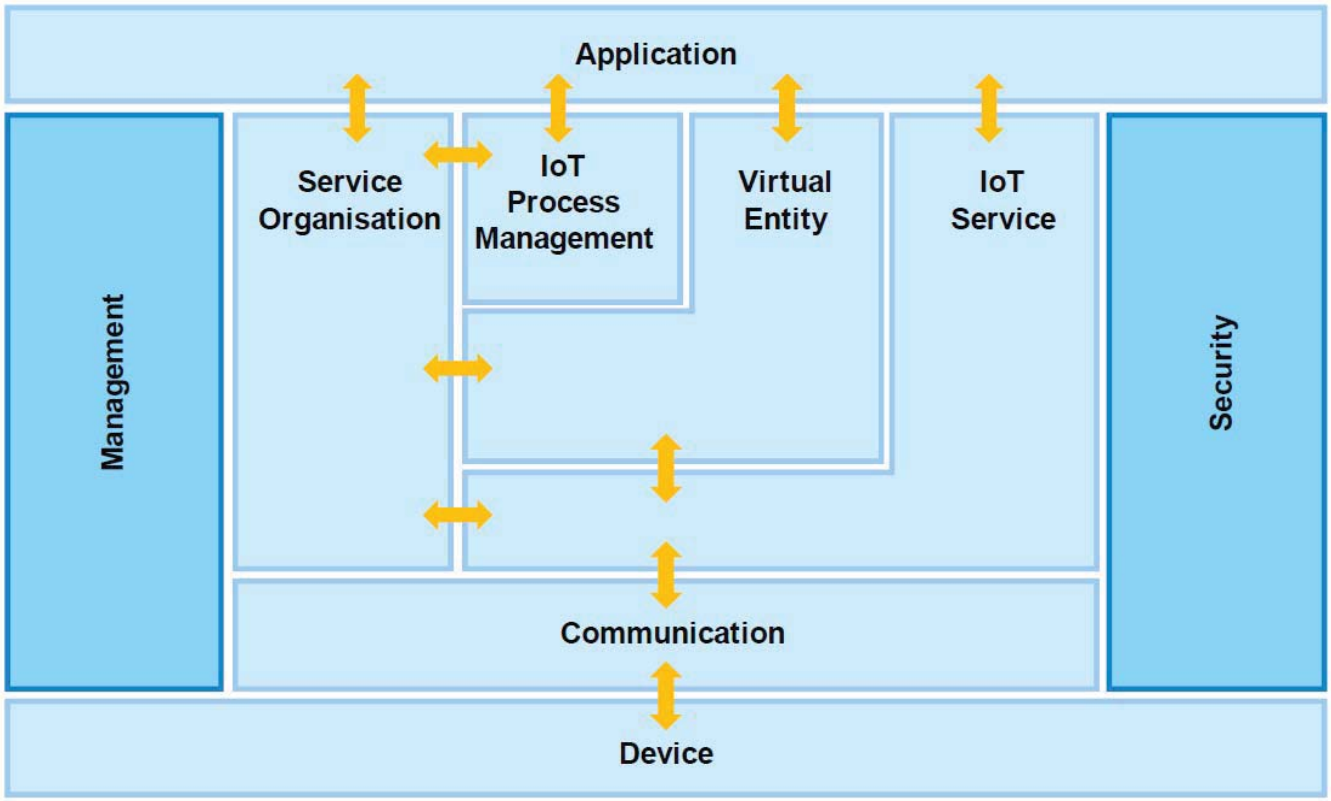


Figure 2 – IoT Functional Model [6]

2.3. Cyber Physical Systems and Lifecycle Management

The CPS Engineering Facet, as proposed by the CPS – Working Group [5], depicted in Figure 3, focuses on how CPS are made, using layers typical for engineered systems, such as Business, Lifecycle, Operation and Physical. The Business Layer represents societal, business and individual Requirements that needs business enterprises response. Existing and emerging government Regulation is another important part of the Business Layer. For large distributed CPS with many conflicting operational objectives, Incentives are important tools for coupling the business layer to all phases of CPS life cycle. The CPS lifecycle, similar to other engineered products, covers phases from engineering design through manufacture, to operation and to disposal of products. The Life Cycle Management Layer represents the four phases of CPS lifecycle. The Operations Layer extends to functionalities and services implemented by the networked interaction of cyber and physical components. The role of Cyber-Physical Abstraction Layers is to ensure that essential properties (such as stability or timing) are guaranteed by the introduced invariants. Among the many abstractions that are applied to CPS, functional abstractions are of special interest. The functional abstraction describes how a CPS is logically decomposed into components and a structure in which these components relate to and interact with each other to form the full system functions. Finally, the Physical Layer represents the physical part of CPS. All CPS incorporate physical systems and interactions implementing some forms of energy and material transfer processes. Physical systems include plants, computation and communication platforms, devices and equipment [5].

Phases of the Lifecycle Management Layer

Design: *Current engineering design flows are clustered into isolated, discipline-specific verticals, such as CAD, thermal, fluid, electrical, electronic control and others. Heterogeneity and cross-cutting design concerns motivate the need for establishing horizontal integration layers in CPS design flows. This need can be answered only with the development of new standards enabling model and tool integration across traditionally isolated design disciplines.*

Manufacturing: *CPS manufacturing incorporates both physical and cyber components as well as their integration. As product complexity is increasingly migrating toward software components, industries with dominantly physical product lines need to change. This transformation is frequently disruptive, requires the adoption of new manufacturing platforms, design methods, tools and tighter integration of product and manufacturing process design.*

Operations: *CPS operations cover the phase of the life cycle where benefits of new technologies are manifested in terms of better performance, increased autonomy, new services, dependability, evolvability and other characteristics.*

Disposal: *Cost of disposing physical components is integral part of the overall life-cycle management process.*

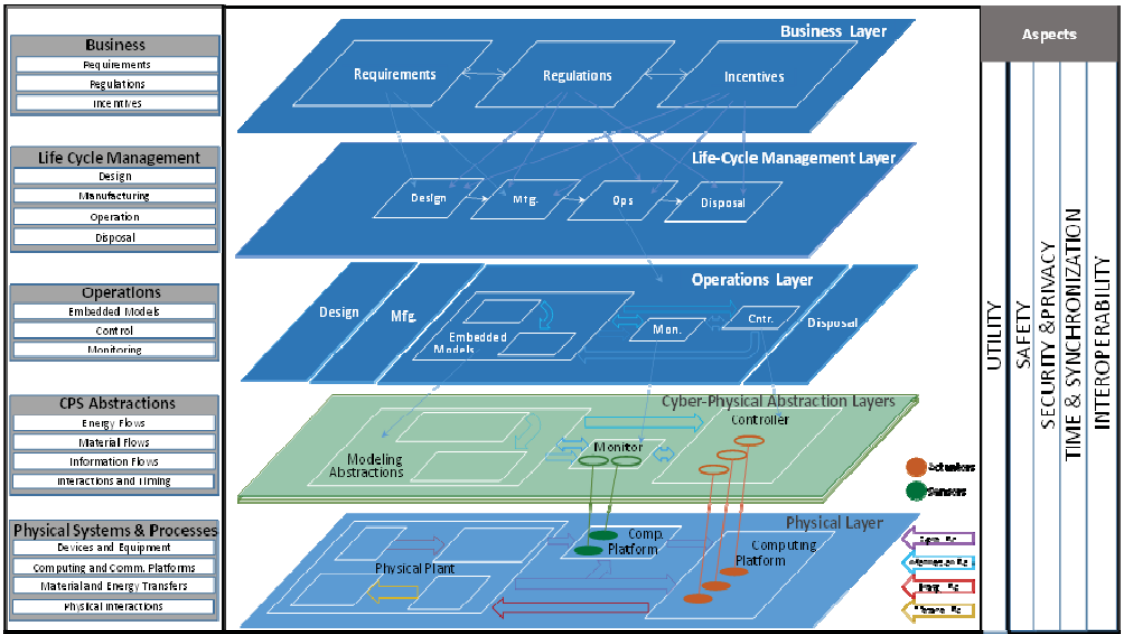


Figure 3 – CPS Engineering Facet [5]

3. Need for a global lifecycle approach

As explained earlier, the role of smart city solutions is becoming bigger in daily city operation. Yet, most of those solutions are vertically locked, where the data collection, processing, analysis and the resulting decisions and accumulated knowledge are normally locked within the boundaries of a particular domain: traffic, parking, energy, water, etc. Although, it is not expected that complete convergence will happen between those verticals; seamless flow of information can help horizontal integration to be realized. Such integration is important for efficiency purposes, taking into consideration that some parts of the value chain are not fiscally feasible or administratively possible to replicate. In this regard, many governments around the world have adopted open data policies to encourage/ oblige government organizations to open up their data, and hence generate economic value and encourage entrepreneurship and innovation.

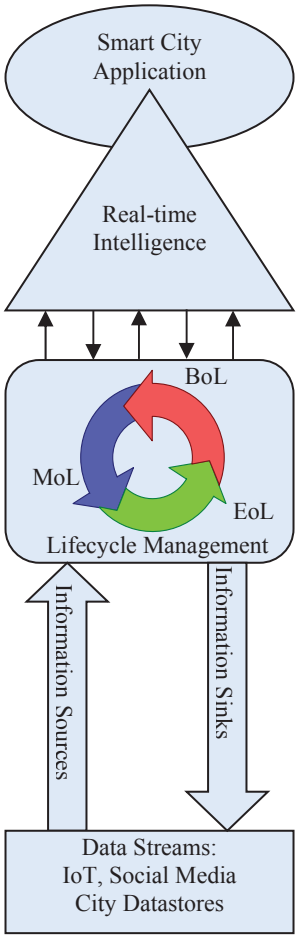


Figure 4: Smart City: High-Level Conceptual Model

On a very high-level, there is a need for a global approach that manages the collected data, processed information and accumulated knowledge according to a lifecycle point of view; and allows seamless flow between different domains, across all phases of lifecycle. To do so, the Smart City Framework (SCF) – discussed in section 2 – could be decomposed in order to decouple the information sources and sinks from real-time intelligence functions. In the meantime, a new Lifecycle Management function could be introduced to manage data, versions, variants and the business processes associated with heterogeneous, uniquely identified connected objects [7][8]. The Lifecycle Management shall support all phases of lifecycle; integrate people, processes, and technologies; and assure information consistency, traceability, and long-term archiving; while enabling intra/ inter-collaboration within the same city and with other cities, if needed [10].

As presented in section 2, the CPS Architecture has proposed lifecycle management layer in its engineering facet. The proposed CPS lifecycle, similar to other engineered products, covers phases from engineering design through manufacture, to operation and to disposal of products. In such a case, Product Lifecycle Management (PLM) has been proven to trace and manage all the activities and flows of data and information during the product development process and also during the actions of maintenance and support [10]. Since the objective of this paper is to consider the entire smart city ecosystem, including stakeholders, systems, processes, etc., Quantum Lifecycle Management (QLM) can be more preferred than PLM. The Open Group has standardized Quantum Lifecycle Management (QLM) as an extension to and derivative of PLM [11]. However, PLM is mainly focused on information about product types and their versions, QLM may be applied to any "object" lifecycle including human, services, applications, etc. [11]. QLM messaging specifications consist of two standards: the QLM Messaging Interface (QLM-MI) that defines what types of interactions between objects are possible and the QLM Data Format (QLM-DF) that defines the structure of the information included in QLM messages [4]. QLM standards can serve the requirements of the smart city high-level conceptual model shown in Figure 3 from different perspectives. The QLM standards, as proposed by The Open Group, provide generic and standardized application-level interfaces [7] in order to create ad hoc and loosely coupled information flows between any kinds of products, devices, computers, users and information systems when and as needed [7]. In addition, QLM applies Closed-Loop Lifecycle Management (CL2M) that enables the information flow to include stakeholders and customers; and enables seamless transformation of information to knowledge [8]. QLM, through CL2M, enhances information security, interoperability, manageability; but most importantly for this research, information visibility and information sustainability to ensure data availability for any system, anywhere, and at any time, while being "consistent" (i.e., not outdated or wrong) [8].

4. Conclusion and Future Work

In this paper, it is proposed to use lifecycle concepts to model the smart city ecosystem. Current smart city, IoT and CPS models are more focused on the engineering system aspect; however, the proposed vision is to consider the entire smart city ecosystem: integrating people, processes, and technologies; and assure information consistency, traceability, and long-term archiving. Although, PLM has been proven very successful to trace and manage all the activities and flows of data and information during the product development process and also during the actions of maintenance and support; QLM adds new capabilities that make it more suitable for smart city modeling.

From another perspective, the proposed approach will develop and promote the smart city ecosystem. Taking into consideration that some parts of the value chain are not fiscally feasible or administratively possible to replicate, the proposed loose-coupling of information from data sources will generate economic value and encourage entrepreneurship and innovation.

However, the presented concepts have shown good level of applicability, it should be subject to more in depth practical test of implementation. The way forward can be using the QLM standards: Data Formats and Messaging Interface to model data exchange between multiple domains in the smart city ecosystem.

5. References

[1] United Nations, Department of Economic and Social Affairs, Population Division. "World Urbanization Prospects: The 2014 Revision, Highlights". (ST/ESA/SER.A/352), 2014.

[2] H. Chourabi, T. Nam, S. Walker, J. Gil-Garcia, S. Mellouli, K. Nahon, T. Pardo, H. Scholl. "Understanding Smart Cities: An Integrative Framework". 45th Hawaii International Conference on System Sciences – IEEE computer society, pp. 2289 – 2297, 2012.

[3] The International Telecommunication Union, ITU-T Focus Group on Smart Sustainable Cities. "Setting the framework for an ICT architecture of a smart sustainable city". May 2015.

[4] The British Standards Institution. "PAS 180:2014, Smart cities – Vocabulary". February 2014.

[5] Cyber Physical Systems Public Working Group. "Framework for Cyber-Physical Systems". Preliminary Discussion Draft, Release 0.7, March 2015.

[6] Internet of Things – Architecture (IoT-A). "Final architectural reference model for the IoT v3.0". http://www.iot-a.eu/public/public-documents/d3.1, 2013.

[7] N. Shrestha, S. Kubler and K. Främling. "Standardized framework for integrating domain-specific applications into the IoT". Aalto University – Finland, 8 pages, 2014.

[8] S. Kubler, K. Främling and W. Derigent. "P2P Data Synchronization for Product Lifecycle Management". Aalto University – Finland, Universit´e de Lorraine – France, 21 pages, 2013.

[9] V. Tsiatsis (editor), et.al. "Real-Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications". EU FP7 CityPulse Deliverable D5.1, 2014.

[10] A. Corallo, M. Latino, M. Lazoi, S. Lettera, M. Marra, and S. Verardi. "Defining Product Lifecycle Management: A Journey across Features, Definitions, and Concepts". ISRN Industrial Engineering, Vol. 2013, Article ID 170812, 10 pages, 2013.

[11] The Open Group QLM Work Group. "An Introduction to Quantum Lifecycle Management (QLM)". November 2012.

Formation of Word Dictionary of Bangla Vowel Ended Roots for First Person for Universal Networking Language

Md. Nawab Yousuf Ali[1], Golam Sorwar[2], and Md. Shamsujjoha[1]
[1]Computer Science & Engineering Dept., East West University, Dhaka, Bangladesh
[2]School of Business and Tourism, Southern Cross University, NSW, Australia
Email: nawab@ewubd.edu, golam.sorwar@scu.edu.au, dishacse@yahoo.com

Abstract - *Interlingua approach plays a vital role in designing a multilingual machine translation system. The Universal Networking Language (UNL) is an international project with an aim to create an interlingua. The motivation behind UNL is to develop an interlingua representation as to semantically equivalent sentences of all languages can have the same interlingua representation. The word dictionary plays an important role to represent native language words in UNL. This paper develops format for word dictionary of Bangla Vowel Ended roots to be incorporated into UNL. The proposed entries are to be used to combine with their inflexions to produce verbs, and hence these verbs can be used for conversion of native language sentences into the UNL expressions. This paper provides the format of vowel ended roots along with their alternatives based on the framework of UNL provided by the UNL center of the Universal Networking Digital Language (UNDL) Foundation.*

Keywords: Dictionary, UNL, Universal Words, Vowel

1 Introduction

The UNL project is a large scale international cooperation with a goal to providing information on the Internet in all national languages of the members of the United Nations [1]. Under this project, a tool, called Enconverter [2], converts each native language sentence into a UNL expression; another tool, called Deconverter [3], translates UNL expression to any native language. The development of language specific components such as dictionary and analysis rules, is carried out by researchers across the world. Dictionary plays a vital role in conversion processes by presenting native languages' words in UNL formats. This paper addresses the following key points associated with the development of format of word dictionary for Bangla vowel ended roots:

- analysis of Bangla vowel ended roots (VER)
- grouping them into categories how verbal inflexions are added with them to form verbs
- finding the alternative roots of the VERs
- outlining the format of VERs, and
- development of dictionary entries of VERs

The rest of the paper is organized as follows. Section 2 describes the structure of UNL. Format of word dictionary of UNL is explained in Section 3. Analysis of Bangla vowel ended roots (VER) is presented in Section 4. This section explains all categorizations of roots with alternative roots of the VERs. In Section 5, we explain the dictionary format of Bangla vowel ended roots and developed dictionary entries of all the VERs along with their attributes. Finally, Section 6 summarizes the paper with some future research plan.

2 Universal Networking Language

The UNL [4] has been defined as a digital meta-language for describing, summarizing, refining, storing, and disseminating information in a machine independent and human language neutral form. It represents information, i.e. meaning, sentence by sentence. Each sentence is represented as a hypergraph, where nodes represent concepts with arcs relationship between the concepts. This hypergraph is also represented as a set of directed binary relations between the pair of concepts present in a sentence. Concepts are represented as character-strings called Universal Words (UWs). Knowledge within a UNL document is expressed in the following three dimensions [4].

2.1 Universal Words (UWs)

Word knowledge is expressed by Universal Words which are language independent. UWs constitute the UNL vocabulary and the syntactic and semantic units which are combined according to the UNL laws in forming UNL expressions. They are tagged using restrictions describing the sense of the word in a current context. For example, *drink*(*icl>liquor*) denotes the noun sense of drink restricting the sense to a type of liquor. Here icl stands for inclusion and forms an is-a relation as in semantic nets [1].

2.2 Relation Labels (RL)

Conceptual knowledge is captured by the relationship between Universal Words (UWs) through a set of UNL relations. For example, *Human affects the environment* is described in the UNL expression as:
{unl}

agt
(affect(icl>do).@present.@entry:01,human(icl>animal).@pl)
obj(affect(icl>do).@present.@entry:01,environment
(icl>abstract
thing).@pl)
{/unl}

where, *agt* and *obj* mean the agent and object respectively. The terms *affect(icl>do)*, *human(icl>animal)* and *environmen(icl>abs-tract thing)* are the UWs denoting concepts.

2.3 Attribute Labels (AL)

Speaker's view, aspect, time of event, etc. are captured by UNL attributes. For instance, in the above example, the attribute *@entry* denotes the main predicate of the sentence, *@present* denotes the present tense, *@pl* is for the plural number and *:01* represents the scope ID. A UNL expression can also be represented as a graph. For example, the UNL expressions and the UNL graph for the sentence, *I went to Malaysia from Bangladesh by aeroplane to attend a conference*, are shown in Fig 1. In the Fig. 1(a), *agt* denotes the agent relation, *obj* the object relation, *plt* the place relation denoting the place to go, *plf* is also a place relation that denotes the place from, *pur* states the purpose relation, whereas *met* is for method relation. U NL expressions provide the *meaning content* of the text. Hence, search could be carried out on the meaning rather than on the text. This, of course,means developing a novel kind of search engine technology. The merit of such a system is that the information in one language can be stored in multiple languages.

{unl}
agt(go (icl>move>do,plt>place,plf>place, agt>thing).
@entry.@past,
i(icl>person))
plt(go (icl>move>do, plt>place, plf>place, agt>thing)
.@entry.@past,
Malaysia (iof>asian_country>thing)) plf(go (icl>move>do,
plt>place, plf>place,
agt>thing) .@entry.@past, Bangladesh (iof> asian_country>
thing))
met(go (icl>move>do, plt>place, plf>place, agt>thing)
.@entry .@past,
aeroplane (icl> heavier-than-air_ craft>thing, equ> airplane))
obj:01 (attend (icl>go_to>do, agt>person, obj>place) .@entry,
conference (icl>meeting>thing) .@indef) pur(go
(icl>move>do, plt>place,
plf>place,
agt>thing) .@entry .@past, :01)
{/unl}

Fig. 1(a): UNL expression of the sentence.

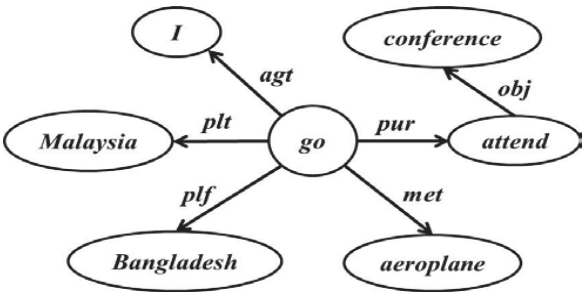


Fig. 1(b): UNL graph of the sentence.

3 Word Dictionary

A Word Dictionary is a collection of word dictionary entries. Each entry of a Word Dictionary is composed of three kinds of elements: the *Headword (HW)*, the *Universal Word (UW)*, and the *Grammatical Attribute (GA)***.** A Headword (*HW*) is a notation/surface of a word of a natural language composing the input sentence and is to be used as a trigger for obtaining equivalent UWs from the Word Dictionary in EnConversion. An UW expresses the meaning of the word and is to be used in creating UNL networks (UNL expressions) of output. GAs are the information on how the word behaves in a sentence and they are to be used in enconversion rules. Each Dictionary entry has the following format of any native language word [5, 6].
Data Format:
[HW]{ID}"UW"(Attribute1, Attribute2,...)<FLG, FRE, PRI>
Here,
HW ← Head Word (Bangla word)
ID ← Identification of Head Word (omitable)
UW ← Universal Word
ATTRIBUTE ← Attribute of the HW
FLG ← Language Flag
FRE ← Frequency of Head Word
PRI ← Priority of Head Word

Elements of Bangla-UNL Dictionary format are shown in Fig. 2.

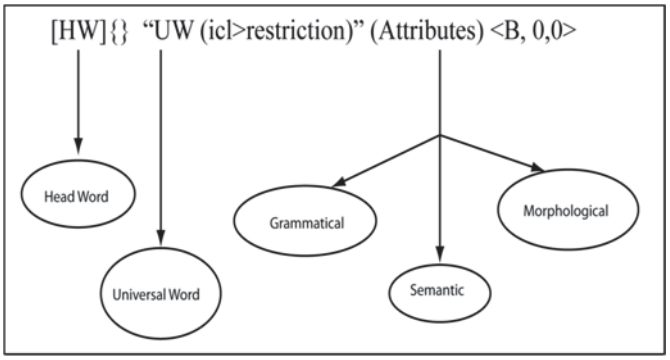


Fig. 2: Format of a Bangla Word Dictionary.

Table 1: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG1 to VEG4 for First Person.

	Tenses	পা (pa)	খা (kha)	গা (ga)	চা (cha)	ছা (ccha)	নি (ni)	দি (ni)	যা (ja)
Present	Pres. Indef.		ই	ই	ই	ই	ই	ই	ই
	Pres. Cont.	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি
	Pres. Perf.	পা>পে য়েছি	খা>খে য়েছি	গা>গে য়েছি	চা>চে য়েছি	ছা>ছে য়েছি	য়েছি	য়েছি	যা>গি য়েছি
Past	Past Indef.	পা>পে লাম	খা>খে লাম	গা>গাই লাম	চা>চাই লাম	ছা>ছাই লাম	লাম	লাম	যা>গে লাম
	Past Habit.	পা>পে তাম	খা>খে তাম	গা>গাই তাম	চা>চাই তাম	ছা>ছাই তাম	তাম	তাম	যা>যে তাম
	Past Cont.	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম
	Past Perfect	পা>পে য়েছিলাম	খা>খে য়েছিলাম	গা>গে য়েছিলাম	চা>চে য়েছিলাম	ছা>ছে য়েছিলাম	য়েছিলাম	য়েছিলাম	যা>গি য়েছিলাম
Future	Fut. Indef.	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব
		VEG1		VEG2			VEG3		VEG4

Some examples of dictionary entries of Bangla language are as follows:
[আপনাকে]{} "you(icl>person)" (PRON, HPRON, HON,SG,P2)
[আপনাদি গকে]{} "you(icl>person)" (PRON,HPRON,HON,PL,P2,,SHD)
[ওরা]{} "they(icl>person)" (PRON, HPRON, PL,GEN, P3,CHL)
[তাঁরা]{} "they(icl>person)" (PRON, HPRON, PL,HON, P3, CHL)
[তুই]{} "you(icl>person)" (PRON, HPRON, SG,NEG, P2)

where, attributes PRON for pronoun, HPRON for human pronoun, GEN for general, NEG for neglect, HON for respect, SG for singular, PL for plural, CHL for conversation language, SHD for literature language, P1 for first person, P2 for second person and P3 for third person respectively.

4 Analysis of Bangla Vowel Ended Roots

For appropriate morphological analysis and design verb root templates, verb roots have been divided into two broad categories: Vowel Ended Group (VEG) and Consonant Ended Group (CEG), according to tenses and persons. Each of them is again divided into sub-groups. Our paper focuses on only vowel ended groups. In Bangla language, 25 vowel ended roots have been found so far [7-10]. After analyzing these roots we have categorized them into 10 groups based on how verbal inflexions are added with them to form verbs. During this categorization, we have considered the behavior of verbal inflexions with first person and tenses (present, past and future). For example: আমি বিশ্ববিদ্যালয়ে যাই, *aami bishabiddaloye jai* means "I go to university". Here, verb is 'যাই', *jai*. In this verb, root 'যা' is a vowel ended root and 'ই' is verbal inflexion. If we write the sentence in present continuous form, we get, আমি বিশ্ববিদ্যালয়ে যাচ্ছি, *aami bishabiddaloye jachhi* means "I am going to university". Although the root is the same as previous sentence but due to a change in tense, the verbal inflexion of the verb is 'চ্ছি'. Whereas, the present perfect form of this sentence is, আমি বিশ্ববিদ্যালয়ে গিয়েছি, *Ammi bishabiddaloye giechhi* means "I have gone to university" . In this sentence, the original root 'যা' is changing its form to 'গি ', *gi*' for making verb গিয়েছি , '*giechhi*', where য়েছি , '*echhi*' is the verbal inflexion. Similar changes have been observed in different roots for different tenses. Tables 1, 2, and 3 present the subgroups of VEG1, VEG2, VEG3, VEG4, VEG5, VEG6, VEG7, VEG8, VEG9, VEG10 and VEG11 along with their inflexions respectively. The tables show the roots with their corresponding tenses for first person. In Table 1, roots পা (pa) and খা (kha) fall into VEG1. They do not change in present indefinite, present continuous, past continuous and future indefinite tenses. But they change themselves from পা (pa) to পে (pe) and খা (kha) to খে (khe) in other tenses. Similarly, roots গা (ga), চা (che), and ছা (chha) in VEG2 are changed to গে (ge), চে (che), and ছে (chhe) in present perfect and past perfect tenses respectively. Roots নি (ni), and দি (di) of VEG3 remain unchanged in all tenses, whereas root যা (ja) in VEG4 is changed to গি (gi) in present perfect and past perfect tenses, গে (ge) in past indefinite and যে (je) in past habitual tenses respectively.

Table 2: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG5 to VEG8 for First Person.

		Vowel Ended Roots								
	Tenses	ছুঁ (chhu)	থু (thu)	শু (shu)	ধু (dhu)	ন (no)	দু (du)	নু (nu)	রু (ru)	ল (lo)
Present	Pres. Indef.	ই	ই	ই	ই	ই	ই	ই	ই	ই
	Pres. Cont.	ছি	ছি	ছি	ছি		ছি	ছি	ছি	ছি
	Pres. Perf.	য়েছি	য়েছি	য়েছি	য়েছি		য়েছি	য়েছি	য়েছি	য়েছি
Past	Past Indef.	লাম	লাম	লাম	লাম		লাম	লাম	লাম	লাম
	Past Cont.	ছিলাম	ছিলাম	ছিলাম	ছিলাম		ছিলাম	ছিলাম	ছিলাম	ছিলাম
	Past Perfect	য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম		য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম
Future	Fut. Indef.	ব	ব	ব	বো, ব		বো, ব	বো, ব	বো, ব	বো, ব
		VEG5				VEG6	VEG7			VEG8

Table 3: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG9 to VEG11 for First Person.

		Vowel Ended Roots						
	Tenses	**হ (ha)**	**ধা (dha)**	**না (na)**	**বা (ba)**	**ক (ko)**	**ব (bo)**	**র (ro)**
Present	Pres Indef.	ই	ই	ই	ই	ই	ই	ই
	Pres Cont.	ছি	ছি	ছি	ছি	ছি	ছি	ছি
	Present Perfect	য়েছি	ধা>ধে য়েছি	না>নে য়েছি	বা>বে য়েছি	য়েছি	য়েছি	য়েছি
Past	Past Indefinite	লাম	ধা>ধাই লাম	না>নাই লাম	বা>বাই লাম	ক>কই লাম	ব>বই লাম	র> লাম
	Past Habitual	তাম	ধা>ধাই তাম	না>নাই তাম	বা>বাই তাম	ক>কই তাম	ব>বই তাম	র>রই তাম
	Past Cont.	ছিলাম	ছিলাম	ছিলাম	ছিলাম	ছিলাম	ছিলাম	ছিলাম
	Past Perfect	য়েছিলাম	ধা>ধে য়েছিলাম	না>নে য়েছিলাম	বা>বে য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম
Future	Fut. Indef.	ব	ব	ব	ব	বো, ব	বো, ব	বো, ব
		VEG9	VEG10			VEG11		

In Table 2, roots ছুঁ (cchu) , থু (thu) , শু (shu) , ধু (dhu) , ন (no) , দু (du), নু (nu) , রু (ru) and ল (lo) of VEG5, VEG6, VEG7 and VEG8 remain unchanged in all tenses. In Table 3, roots **ধা** (dha), **না** (na) and **বা** (ba) in VEG10 are changing to ধে (dhe), নে (ne) and বে (be) in present perfect and past perfect tenses and to ধাই (dhai), নাই (nai) and বাই (bai) in past indefinite and past habitual tenses respectively. And roots ক (ko), ব (bo), র (ro) and ল (lo) in VEG11 change themselves to কই (koi), বই (boi), রই (roi) and লই (loi) in past indefinite and past habitual tenses respectively.

5 Development of Format for Word Dictionary of Bangla Vowel Ended Roots

After detailed analysis of the Bangla vowel ended roots, following template has been developed based on the format in Section 3. We have also meticulously considered the different roots and their alternatives along with their inflexions in the Tables 1, 2 and 3 for generating templates.
[HW]{}"UW(icl/iof…>concept1>concept2…,REL1>…,REL2>…," (ROOT, VEND, DEF/ [,ALT1 / ALT2/ALT3..], VEGn, #REL1, #REL2, … <FLG, FRE, PRI>

where, HW← Head Word (Bangla Word; in this case it is Bangla root);
UW← Universal Word (English word from knowledge base);
icl/iof/… means *inclusion/instance of* …to represent the concept of universal word
REL1/REL2.., indicates the related relations regarding the corresponding word.
ROOT ← It is an attribute for Bangla roots. This attribute is immutable for all Bangla roots.
VEND ←is the attribute for vowel ended roots.
VEGn ← attribute for the group number of vowel ended roots (n=1, 2…11).
CEGn ← attribute for the group number of consonant ended roots (n=1, 2…11).
ALT1, ALT2, ALT3 etc. are the attributes for the first, second and third alternatives of the vowel ended roots respectively.
DEF← attribute for default root.
#REF1, #REF2 etc. are the possible corresponding relations regarding the root word.
Here, attributes, ROOT, VEND are fixed for all Bangla roots whereas, ALT1 or ALT2 etc. is not necessary for all roots because they are used only for alternative roots.
In the following examples, we are constructing the dictionary entries for some sample verb roots using our designed template:
[যা]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, VEG3, #PLF, #PLT, #AGT)<B, 0, 0>
[গি]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, ALT, VEG3, #PLF, #PLT, #AGT) <B,0,0>
[খা]{}"eat(icl>consume>do,agt>living_thing, ins>thing, obj>concrete_thing, plf>thing, tim>abstract_thing)" (ROOT, VEND, VEG1, #PLF, #PLT, #AGT)<B, 0, 0>
For first two entries the relation *plf* (place from) indicates from where agent go/goes, plt (place to) means to where go/goes, agt (agent) for who go/goes and attribute ALT1 indicates that root "গি" (*gi*) is the first alternative of root "যা" (*ja*) shown in Table 1. Attributes #PLF, #PLT and #AGT indicate that relations *plf, plt* and *agt* can be made with roots "গি" (*gi*) and "যা" (*ja*). Similarly, other entries have been developed according to the format discussed above. Our Proposed Dictionary Entries of vowel ended roots along with their alternatives are given below.

- Dictionary Entries of VEG1:

[পা]{}"get((icl>do,equ>obtain,src>uw,agt>thing,obj>thing)" (ROOT, VEND, DEF, VEG1,#OBJ,#AGT)<B, 0, 0>
[পে]{}"get((icl>do,equ>obtain,src>uw,agt>thing,obj>thing)" (ROOT, VEND, ALT1, VEG1,#OBJ,#AGT)<B, 0, 0>
[খা]{}"eat(icl>consume>do,agt>living_thing,obj>concrete_thing,ins>thing)"
(ROOT, VEND,DEF,VEG1,#AGT,#OBJ,#INS)<B,0,0>
[খে]{}"eat(icl>consume>do,agt>living_thing,obj>concrete_thing,ins>thing)" (ROOT, VEND,ALT1,VEG1,#AGT,#OBJ,#INS)<B,0,0>

- Dictionary Entries of VEG2:

[গা]{}"sing(icl>do,com>music,cob>thing,agt>living_thing, obj>song,rec>living_thing)" (ROOT, VEND, DEF,VEG2, #AGT,#OBJ,#COM,#COB,#REC) <B, 0, 0>
[গে]{}"sing(icl>do,com>music,cob>thing,agt>living_thing, obj>song,rec>living_thing)" (ROOT, VEND, ALT1,VEG2, #AGT,#OBJ,#COM,#COB,#REC) <B, 0, 0>
[গাই]{}"sing(icl>do,com>music,cob>thing,agt>living_thing ,obj>song,rec>living_thing)" (ROOT, VEND, ALT2,VEG2, #AGT,#OBJ,#COM,#COB,#REC) <B, 0, 0>
[চা]{}"want(icl>desire>be,obj>uw,aoj>volitional_thing,pur >thing)"(ROOT,VEND,DEF, VEG2, #OBJ,#AOJ,#PUR)<B,0,0>
[চে]{}"want(icl>desire>be,obj>uw,aoj>volitional_thing,pur >thing)"(ROOT,VEND, ALT1, VEG2, #OBJ,#AOJ,#PUR)<B,0,0>
[চাই]{}"want(icl>desire>be,obj>uw,aoj>volitional_thing,pu r>thing)"(ROOT,VEND, ALT2, VEG2, #OBJ,#AOJ,#PUR)<B,0,0>
[ছা]{}"roof(icl>cover>do,agt>volitional_thing,obj>thing,in s>thing)" (ROOT, VEND, DEF, VEG2, #AGT,#OBJ,#INS)<B, 0, 0>
[ছে]{}"roof(icl>cover>do,agt>volitional_thing,obj>thing,in s>thing)" (ROOT, VEND, ALT1, VEG2, #AGT,#OBJ,#INS)<B, 0, 0>
[ছাই]{}"roof(icl>cover>do,agt>volitional_thing,obj>thing,i ns>thing)"(ROOT, VEND, ALT2 VEG2, #AGT,#OBJ,#INS)<B, 0, 0>

- Dictionary Entries of VEG3:

[নি]{}"take(icl>capture>do,agt>thing,obj>thing)"(ROOT, VEND, DEF,VEG3, #AGT, #OBJ)<B,0,0>
[দি]{}"give(icl>do,equ>hand_over,agt>living_thing,obj>co ncrete_thing,rec>person)" (ROOT, VEND, DEF, VEG3, #AGT,#OBJ,#REC)<B, 0, 0>

- Dictionary Entries of VEG4:

[যা]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, VEG4, #PLF, #PLT, #AGT)<B, 0, 0>
[গি]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, VEG4, #PLF, #PLT, #AGT)<B, 0, 0>
[গে]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, VEG4, #PLF, #PLT, #AGT)<B, 0, 0>
[যে]{}"go(icl>move>do, plf>place, plt>place, agt>thing)" (ROOT, VEND, VEG4, #PLF, #PLT, #AGT)<B, 0, 0>

- Dictionary Entries of VEG5:

[ছুঁ]{}"touch(icl>come_in_contact>do,agt>person,obj>conc rete_thing,ins>thing" (ROOT, VEND, DEF,VEG5, #AGT,#OBJ,#INS)<B, 0, 0>
[থু]{}"put(icl>displace>do,plc>thing,agt>thing,obj>thing)"(ROOT,VEND,DEF, VEG5, #AGT,#OBJ,#PLC)<B, 0, 0>
[শু]{}"sleep(icl>rest>be,aoj>living_thing)"(ROOT,VEND, VEG5,#AOJ,#PLC)<B,0,0>

[ধু]{}"wash(icl>serve>do,agt>living_thing,obj>concrete_thing,ins>functional_thing)" (ROOT, VEND, DEF,VEG5,#AGT,#OBJ,#INS)<B,0, 0>

- Dictionary Entries of VEG6:

[ন]{}"be(icl>be>not, aoj>thing)" (ROOT, VEND, DEF, VEG6, #AOJ)<B, 0, 0>

- Dictionary Entries of VEG7:

[দু]{}"milk(icl>draw>do,agt>thing,obj>thing)" (ROOT, VEND, DEF, VEG7, #AGT, #OBJ)<B, 0, 0>
[নু]{}"bath(icl>vessel>thing)" (ROOT, VEND, VEG7, #PLF, #PLT, #AGT)<B, 0, 0>

[রু]{}"sow(icl>put>do,plt>thing,agt>thing,obj>concrete_thing)" (ROOT, VEND, DEF,VEG7,#PLT, #AGT,#OBJ)<B, 0, 0>

- Dictionary Entries of VEG8:

[ল]{}"take(icl>require>be,obj>thing,aoj>thing,ben>person)" (ROOT, VEND, DEF, VEG8, #OBJ, #AOJ, #BEN)<B, 0, 0>

- Dictionary Entries of VEG9:

[হ]{}"be(icl>be,equ>be_located,aoj>thing,plc>uw)" (ROOT, VEND, DEF, VEG9, #AOJ, #PLC)<B, 0, 0>

- Dictionary Entries of VEG10:

[ধা]{}"urge(icl>rede>do,agt>volitional_thing,obj>volitional_thing,gol>thing)" (ROOT, VEND, DEF,VEG10, #AGT,#OBJ,#GOL)<B, 0, 0>
[না]{}"bath(icl>vessel>thing)" (ROOT, VEND, VEG10,#AGT,#PLC)<B,0, 0>
[বা]{}"row(icl>move(icl>cause)>do,plt>thing,agt>person, obj>boat,ins>thing)"(ROOT,VEND, DEF,VEG10, #PLF, #PLT, #AGT,#OBJ,#INS)<B, 0, 0>

- Dictionary Entries of VEG11:

[ক]{}"talk(icl>communicate>do,cob>uw,agt>person,obj>thing,ptn>person)" (ROOT, VEND, DEF, VEG11, #AGT,#OBJ,#PTN,#COB)<B, 0, 0>
[ব]{}"bear(icl>have>be,obj>property,aoj>thing)" (ROOT, VEND, DEF, VEG11, #OBJ, #AOJ)<B, 0, 0>
[র]{}"stay(icl>dwell>be,aoj>person,plc>uw)" (ROOT, VEND, DEF, VEG11, #AOJ, #PLC)<B, 0, 0>

6 Conclusions and Future Work

This paper has analyzed the Bangla vowel ended roots and grouped them into different categories based on how verbal inflexions are added with them to form verbs for first person, and then outlined the format of word dictionary for the roots. In this paper, we have also developed word dictionary entries for all vowel ended roots. These entries can be used to create verbs combining with their respective verbal inflexions. A Bangla native language sentence with verb can be easily converted into UNL expression by analysis rules which can later be converted into any other languages using language specific generation rules. Our future research is to develop formats of Bangla vowel and consonant ended roots for first, second and third person of all tenses. The proposed format is expected to be equally applicable to other languages with verb roots.

7 References

[1] M. E. H. Choudhury, M. N.Y. Ali, M.Z.H. Sarkar, R. Ahsan, "Bridging Bangla to Universal Networking Language- a Human Language Neutral Meta- Language", International Conference on Computer and Information Technology (ICCIT), Dhaka, pp.104-109, 2005.

[2] EnConverter Specification, Version 3.3, UNL Center/UNDL Foundation, Tokyo 150-8304, Japan 2002.

[3] DeConverter Specification, Version 2.7, UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002.

[4] H. Uchida, M. Zhu, and T. C. D. Senta, Universal Networking Language, UNDL Foundation, International environment house, Geneva, Switzerland, 2005/06.

[5] UNDL Foundation: The Universal Networking Language (UNL) specifications version 3.2, 2003.

[6] J. Pairkh, J. Khot, S. Dave, P. Bhattacharyya, "Predicate Preserving Parsing", Department of Computer Science and Engineering, Indian Institute of Technology, Bombay.

[7] H. Bondopoddaye, "Bongioi Shobdokosh", Shahitto Okademy, Calcutta, 2001.

[8] D. M. Shahidullah, "Bangala Vyakaran", Maola Brothers Prokashoni, Dhaka, pp.110-130, August 2003.

[9] D. C. Shniti Kumar,"Vasha-Prokash Bangla Vyakaran", Rupa and Company Prokashoni, Calcutta, pp.170-175, 1999.

[10] D. S. Rameswar, "Shadharan Vasha Biggan and Bangla Vasha", Pustok Biponi Prokashoni, pp.358-377, 1996.

Analyzing the Role of Child Friendly Environments in Continuation of Sustainable Urban Development.

Employing the sense of positive thinking approach

Mohammad Mehdi Khabiri [1], **Amir Reza Khabiri Khatiri**[2]
[1]PHD Candidate/Faculty Member, Architecture & Urban Design Department ,Islamic Azad University of Khormooj, Iran
[2]Architecture & Urban Design Department ,Islamic Azad University of Sirjan, Iran

Abstract- *Cities are the ground for human manufacturing and artifacts and most important of all they are the cradles for living. In recent years the environmental stress and depression in the population especially in urban areas increased notably. Some researchers believe that one of the dozens of reasons for this is to be found in urban architecture .Because the quality of public spaces in different levels affect people's behavior. In urban areas sustainable approach considering enhancing quality of life by improving mental health is the main branch of urban planning. Therefore, the role of city and urban development can't be denied in decreasing the mental health of the citizens and its effect on lie stress and tolerance of citizens. In addition children thought always in statistical analysis a significant population group in urban space and environment users and practically they have the least effect on the parameters of the urban space. Thus the recognition of children as a citizen with equal access to all services is one of the main issues in supplication of urban life. Freedom and flexibility of space, paying attention to the criteria of safety, physical space openness and transparency in the same connection and etc can be either specific or general as a particle means to from environment and urban spaces based on the needs and characteristic of effective learning and development of children but in addition to the development of child friendly environment creation, Understanding and reaching to a positive thinking is special and important factor that we consider to be the missing ring in the chain of continuity, and sustainable urban development can be the basic and supplementary indicators of a child-friendly environment. Therefore, in this study we aimed to examine the role of child upbringing in continuation of the process of sustainable development and defining a new model as well as the physical and visual child friendly environment is created and achieved. In order to achieve this goal the library and analytical methods and the use of technology experts are employed to derive some general conclusions.*

Key words: Sustainable urban development, Citizens mental health, Psychology, child friendly environment, positive sense of neglect.

1. Introduction

Community-environmental based sustainable preservation and economic values alongside with problems of economic development came into existence only after World War II and caused environmental problems, also human class differences were introduced and was the beginning of the return to the nature after a century of dominance of industry on humans' life. The concept of development also covers the aspects of the mental health community in addition to the physical-psychological concepts. In fact with the process of urbanization and urban development of human life, concepts of sustainable development has also been included in the issues of city and urban planning. Entropy (irregularities) is considered as a major barrier for sustainable development which in itself is an organic reaction that appears in urban society. In fact if the behavioral manifestations occur where citizens reduced quality of life, ultimately sustainable development will be an obstacle to movement. Happiness factors for every society depends on time, place and environment. Therefore behavioral abnormalities citizens "stress, intolerance, violence and etc "by any person influences others and makes the society experience grief and sorrow. Providing community with sadness and stress tolerance may be able to control the environment the underlying health of the community and create the enabling environment and prevalence of social development in the city's cultural and behavioral norms which would pave the way for the establishment and organization of cultural cities.

Today urban spaces, unlike their expected definitions, have become unwanted environments so that all of them are considered as unpleasant. The urban environment or Spaces are considered, as the name expresses, spaces used for different age groups and genders and there are many diversities. The urban context of social life in every period of several factors that led to their formation. But what is sustainability conservation and preservation of environmental quality of urban spaces as appropriate the interactions and behaviors helps citizens and caused the formation of a living organism and social dynamics. Aside from the aesthetic qualities of urban spaces continuity and quality of the environment In order to provide comfort and physical and mental health of its citizens. generally, the quality of urban spaces certainly help to adapt the psychological effects of

urban environment, meaning that the predicated performance in sustainable urban development, thereby sustaining the quality of the urban environment for the exchange of culture, entertainment, pleasure and leisure, social life and the exchange of ideas, perspectives, the opinions and views for all sexes and ages for all citizens.

In order to achieve the patterns and concepts of sustainable development in cities, in addition to considering the current time, future of the city and citizens are important and perhaps paying attention to the future in order to continue the use of resources according to the definitions of urban sustainable development is more important. Hence, education and training, in order to take advantage of mutual interest between urban space and the citizens is important. Based on psychological theories of Freud all social anomalies of seniors and citizens are rooted in suppressed instinct during the childhood. This instinct and appropriate or inappropriate responses to them depends on environmental conditions and psychological aspects governing the environment and also cultural conditions. So in order to investigate and study the practical impact of the environment on behaviors and activities of people in urban areas and increasing the satisfaction of citizens from the city, which is the main subject of environmental Psychology in the field of architecture and urbanism, issues of education and training are considered as a dependent variable to psychological factors such as self-esteem and sense of positive thinking.

2. The importance of childhood in continuation of the process of sustainable development

2.1. Environmental Psychology of a Kid Friendly Space

The importance of this issue appears when the environmental psychology is used in order to create healthy interactions and projection of normal behaviors in urban environments. But what is taken from psychological research is that impulsiveness is a function of the environments condition in which it occurs and the users' behavior is affected by the environment at different levels. The purpose can be seen in the different personalities of people of different races due to the nature of the environment where they live. The goal of environmental psychology is better environmental management in order to better mental development and enhancing and strengthening thebehavioral indicators right of citizens living in the present and future so that a particular approach to recover the quality of life and the continuation of the citizen will be helped in future .This group of psychologists believe that the design of the environment have to maximize the freedom of motion and flexibility to the extent that in addition to that environment and its surroundings it will be functional. And psychologically answers to the indicators and standard which are along with the standards. In order to achieve this end and approaching a pattern that covers certain civilian addressees (children), we will consider the concept of realm one of the factors that affect the behavior of addressees. It is possible to say that the depth of the space used by vulnerable groups in terms of sexuality and age are among those parameters that reduce or increase the amount and the way of manipulating people's psychological part of life. Each realm can include big or small social units, groups and systems and make them a healthy place to manifest social behaviors and relations. Many social behaviors have realm expanding sides that should be defined in urban spaces and change identity according to different places, so we can consider the way of pointing out the quality of social relations that bestow identity, social supervising and public participation among the social factors that are affected by the quality and the quantity of the realm and several other anatomic and functional factors can be analyzed in terms of organizing the visual factors and balancing the open and close areas and defining the hierarchy of categorizing in the framework of hierarchy of performance and the hierarchy of accessibility (Bahreini 1378, 28). But the factors that can affect the concept of realm more than others are the mental and psychological factors that appear through time and because of human (citizen) as user of space get a certain importance. By the way, this topic can be examined in two ways. First in hierarchy of human needs the issue of security and comfort have specific importance that can be analyzed and examined in any public or private area and include the sense of security, the sense of belonging, readability of space, satisfaction of needs and motives of the presence in the area and the compatibility of them with the urban performance of the citizens, and on the other hand considering the definition of sustainable development "time" is a factor that influences the main and marginal factors in a way that eventually affecting the perceptional and psychological factors.

2.2. Kid Friendly Space and Sustainable Urban Development

Although children are among the considerable population groups in categories of urban space users, practically they have the least effect on factors shaping the urban space. In this way, besides mentioned factors for designing sustainable urban spaces the factors of education and training in time periods to achieve sustainable development concepts and patterns gain a considerable amount of importance. In the meantime analyzing the existing designed spaces for children we can realize those areal shortcomings and wrong patterns that are incompatible with children's behavioral factors in urban spaces that they spend most of their time in, produce unorthodox form of behavior and social discrepancies in adults. However, some of these anatomic and functional spaces may have purposeful trainings and patterns, but because they are not in line with children's feelings they won't be able to respond to their adolescent needs and in a greater scale the social discrepancies. Children want to fit in and mingle with those who are in the same age with them, they also like to spend their time in natural areas and be active in those places, accordingly the psychologists who work with places that are agreeable to children believe

that there is a bilateral relation between children and the places. In the most ideal mental situation they tend to design their environment in accordance to their own behavioral priorities. In this way the compatibility of the environment along with the flexibility of space and the distinction of the suitable realm and environment for children, prepare the psychological aspects of the environment for them as much as possible.

Every man's character is shaped in his childhood and the way he is brought up has a decisive influence on his feelings and sentiments, thus recognizing children as equal citizens who have accessibility to all kinds facilities is one of the important instances in the urban life and creation of environments agreeable to children (Tosei 1369: 18). According to the definition public urban spaces are the telling tongue of the cities. A child loving city is not only a place in which a child's basic rights like health, transportation, support and education are fulfilled, but also it includes the possibility and competence through which the children are able to create spaces and environments suitable in conveyance of behaviors and activities in accordance with foretold functions in terms of physical and mental aspects, so they are able to communicate and cooperate with the society and in some way they participate in shaping the urban environment (Riggo: 2002: 46).

Since the approval of children's rights treaty several important actions started in order to specify children's rights more than before and recognize them as the main part of society which is influential in society's sustainable development, and encourages them to make creative and affective decisions. The first step in this field is paying attention to their daily life and understanding their perception of the environment. From this point of view, there is a close relation between children's psychological issues and stability that can help solving some challenges of sustainable development. According to these different definitions and several announcements presented in international conferences regarding sustainable development, it is the consequence of growth and awareness of the societies resulting from universal issues and environmental problems like social and economic issues, inequality and most important of all the concerns regarding a healthy future for human being. Since the attempt for achieving a time or place related situation in which the quality of progress and suitable unfettered development in an unending period of time relying on keeping the environmental values tends to be fruitful. There is no doubt the value and place of personal and psychological education of citizens is specified. Regarding the necessity of this issue in international conferences the universal organizations emphasize the issue and the stress is on effectiveness of these issues in cultural and social institutes, and also their application by the young generation makes them sensitive mentally and physically to the sustainable viewpoints.

Goleman in 1995 suggests that nowadays children are more exposed to problems mentally compared to previous generations. They are lonelier, angrier, more depressed, also unrulier and more apt to get worried than before. Many behavioral experts and children's psychiatrists believe that the role of schools and educational centers is pivotal and the solution is to build up a relationship between families and schools, but the thing that environmental psychologists agree upon is that the living environment and urban space are two factors that are more influential and important. John Duty in his book " how to think" suggests that the class is a place where the student gets to know about mental skills and mental conditions and in order to achieve them and also to know about the proper social conditions he needs to learn them in classroom, but it is the urban environment which transfers and maintains these learnings in social dimensions and prepares the ground for their behavioral continuation and enhancing the procedure of sustainable development.

3. Accelerating Sustainable Development by Emphasizing on Structure of Kid Friendly Environments in Urban Spaces

Because of human needs in process of designing and creating the space and also importance of users' behaviors in achieving the goals of sustainable development and moreover regarding the childhood as the most important period in one's life in terms of upbringing, education, designing environments for children according to sustainable development it is necessary to include the grounds of learning and education in the process. Regarding this issue studying children's environments as the places for conveying their sudden behavior based on their sudden physical and mental needs many studies are performed all based on theories about children's growth and learning, in order to shape their environments and the results are presented. This type of psychological approach in the realm of urban architecture and design started in 18th Century, but at ending decade of 20th century and the integration of psychological theories the creation of places and environments for children gained more attention, and researchers and designers through joining these two trends got much success in the field of environment and environmental psychology, and also in creating the environments friendly to children. This made the old structure change significantly and shaped many new approaches based on the synchronic relation of learning and education and creation of spaces in line with increasing children's role in unlimited periods of time in society. Since these approaches on one hand made conformity in psychological dimensions of children's behavior to the functional dimension of places on the other hand with the pattern of participating citizens in the process of sustainable development. As the main frame of the spaces design and child-friendly environment and sustainable urban development was proposed and accepted. The effects of approaches, in addition to creating pleasant feeling of mental and functional aspects of its compliance activities and interaction with the people, increased the quantity and quality of teaching-learning and training the next generation to sustain and reinforce the norms of behavior and citizen participation in governance and management of the city and ultimately strengthen citizen’s sense of belonging to the

cities. Divers and comprehensive strategies and models are presented in research studies design of urban spaces based on the need of children, which are deduced based on different approaches:

Table 1- Principles and design criteria

Provide suitable locations for urban educational facilities for children in neighborhood .	Securing the educational spaces designed to create peace in the neighborhood of educational facilities for children.	Separation of the neighborhood way: riding and walking routes (especial routes for walk)
		Keeping the spaces for children in sight and aroid hiding them.
		Providing spaces for neighborhood residents sit near or adjacent the educational facilities which are designed to create indirect supervision of adults.
		Embedding this spaces as possible along the adjacent residential neighborhood as that these spaces are placed in front of the houses windows to let the inhabitants of the houses, that are mostly parent, monitor the children's behavior and security indirectly.
		Determine the scope of experimental spaces for children's play and learning experiences for children with determination privacy.
		Designing the spaces away from places where is most commonly used by strangers (people who they don't belong to neighborhood)
	Relaxation spaces designed	Designing blocking edges in many places to limit road way.
		Relaxing the designed area using the design and placement of trees, lawns, shrubs and flowers around it.
		Removing the road way in order to relax these places and considering the designed area in the farthest places of sound generator in the neighborhood.
Given the characteristics of the identity of the neighborhood, especially children.		Considering the children's psychology to design the educational facilities.
		Considering the specialists thoughts about child and urban spaces and exerting them in neighborhood design.
		Considering the favorite children's entertainment: cultural, local and ethnic games.
Basie needs of children in the use of educational facilities in the neighborhood.		Considering the morphology and children standards in the design of educational spaces and urban furniture.
		Considering the disabled children and designing the spaces in a way that these children would be able to use them.
		Considering the educational facilities for all neighborhood residents, specially poor children.
Flexible learning facilities and their furniture.		Flexing the educational urban furniture in neighborhood for different age groups use, using a design that has the ability to change size.
		Flexing the spaces and urban furniture by giving various functions to them.

Table 2-design approaches based on children's needs in the general neighborhood.

Spatial-visual criteria	Socialnorms (criteria)	Environmental criteria
. Relaxing the local traffic using impediments in appropriate places or reducing the width of the roadway. . Restrict the movement of vehicles by making twist and turns in the path. . Using light and cheerful colors and avoiding high contrast. . Applying signs in the neighborhood for children in order to orientation and sense of belonging to the neighborhood. . Designing flexible spaces in such a way that there is a possibility of changing a layout of space by children. (In order to promote creativity in children) .Creating temporary exhibitions and workshops for children to encourage pause in space and promoting social interaction . Furniture placement in varied and appropriate to the scale and needs of children in parks and public areas for their convenience and creativity (considering both boys and girls and children with disabilities) . Designing spaces that may have search and discovery ability and could be an exciting space for children .	. Applying proper lighting in the environment when children attend for work to increase the security in the environment. . Survey of children and their participation in the construction of spaces required for this age group. . Safe spaces for children's play and activity of commuter transportation. . Play spaces for children away from strangers and non-local traffic.(away from non-local users and access) . Providing spaces for parents site near the children's playground for social monitoring of indirect. . Spaces should be designed and applied when the sight of the residents exist. . Established a comfortable and independent for children to the neighborhoods social areas due to the security and relaxation.	. Use single leafy trees for shading and climate comfort in social neighborhood areas. . Taking advantage of natural elements in walls, floors, as well as the construction of children play equipment in social neighborhood areas. . Using sand and water play areas for children to provide spiritual needs of children to nature. . Children's play areas should be away from the crowded and blatantly to reduce environmental pollution and preserve peace. . Creating different spaces with a change in slope of the land and vegetation as well as the incorporation of natural elements.

Pay attention to freedom and flexibility of space, performance indicators with regard to visual markings, criteria of safety and physical emotional space while accountability and transparency of information in the area and orientation in space, using exciting colors, space group approaches, creating an evocative sense of space, participating in creating and maintaining the continuity of space and etc. which all can be used special or general as practical ways to shape the environment and urban space according to the needs and characteristics affecting children's development and learning but in addition to the above which build the formation of a child-friendly environment, understanding and achieving the feeling that you are a good person and you are known such between others, is a very

important and special factor. And we believe that as the missing link in the chain of sustainable urban development is a continuing basis and complementary indicators of child-friendly environment.

The fact of everything comes from human mind, which means humans are made from their thoughts and the quality of our mind helps us to be successful and power full or vice versa and in fact, positive thinking helps us to make beneficial changes in behavior, speech, work and our life. positive psychology is a new field in behavioral sciences. This field strings a good view of life to happiness and mental health in personal or social scales.

Martin Sigman believes that positive psychology has an effective and undeniable contribution in growth and development of individuals, families and society.

In today s' world feelings and emotions are disappeared because of secondary factors of life and making social relationships, which is a structural and principles of alive and vibrant communities, is forgotten due to the social wrongs and failures and children and teenagers are directly or indirectly affected by mental and behavioral failures. Psychologists be live that it is possible to interpret and change events and exteriors to opportunity according to strengthen and faith of people using Dynamic urban environments which provide the communication and interaction of citizens. Actually making an environment can play a role in having a positive feeling experience about children, here are the methods. Fist is welcoming happiness and psychological support of the child and the second stage is increasing opportunity for children to use green spaces and natural areas and enjoying them. These child behavioral domains psychologists believe that both immediate pleasure and lasting satisfactions are for the formation and strengthening of negligence in a positive sense, especially children.

Hence they studied samples and found that smooth and trouble-free with the social environment in urban areas leads to a set of the desires. Satisfaction, optimism and behavior as normal and healthy communication. Therefore we can observe ,feel prove a direct interaction between the effects of positive thinking and social life with active areas which have provided the participation and involvement of citizens. Which means although at first sight and start point positive thinking is found embodiment in one, the result of the collective experiences and processes will be a form of social interaction. But at its core reciprocating action and reaction, Leading to the creation of public space in the city in their ideal meaning. So specialist believe that Dynamic Behavior of urban space as a lining organ and sociology can cause the creation development and strengthening the background and context of learning, psychological and behavioral indicators of child friendly cities with a sustainable development approach, through enhancing the sense of positive thinking but if you have designed and created the structural factors of the environment according to what they said.

Result:

The conclusion include that implicitly or comprehensive product and can be compared with what has already been done, generally, most of urban design respond to the constraints defined by the very economic, physical and cultural best deals and the psychological needs of the citizens in the life of the project fail. Predictions of development projects are often based on functional requirements and are indifferent to the needs which define their time not only in the area but also in psychological and sociological context of the plan. Paying attention to the indicators of child – friendly and based on the psychological characteristics of children in the creation of urban space, especially in project that aim to engage citizens specially children and is in sync with their needs and demands, in addition to providing a safe environment with creating a sense of participation in children though reinforcing their positive sense of neglect, will be educating a generation by raising awareness and creating and strengthening the social and creative participation in society, guarantee sustainable development of urban mobility, nature has always been a source of peace, joy, happiness, inner peace is the drive mechanism due to its beautiful and pleasant elements. Enlargement of the city and the development of purely economic has risked the relation of city and citizens with nature. However, architects, planners and urban planners can use the nature-oriented models to reinforce the positive feeling in citizens.

According to the current situations we believe that with creating spaces for children in different levels of behavioral and psychological, positive thinking will be strengthened in the society. Awareness, motivation and responsibility has made the citizens sensitive about the environment culture and the future of the cities. And in future we will have citizens with satisfaction and identification along the healthy mental and emotional participation in sustainable development policies, which they know them as their task of unconscious.

5 References

[1] Altman, Iron, " Environment and Social Behavioral" , Namaziyan,Ali, Tehran, Shahid Beheshti University Publication, 2003

[2] Pakzad, Jahanshah, "Environment Psychology Alphabet for designers", Tehran, Armanshar Publication, 2002

[3] Shieea, Ismaeil, "Preparing City for Kids ", Tehran, Shahr Publication, 1937

[4] Brek, Laura, "Psychology of Growth", Seyyed Mohammadi, Yahya, Tehran, Arasbaran Publication, 2002

[5] Shibani, Mehdi, "Kid's Foot print in Urban Scape", Tehran, Urban Design Topics Publications, Vol 34.

[6] Matin, Cliff & Shirley, Peter , "Urban Space Design Based on Sustainable Development", Molla Yousef, Narsis, Tehran, Payam Publications, 2008

[7] Leng, John, "Creating Architecture Theory: Role of behavioral Science in Environment Design", Eyni Far, Alireza, Tehran, University of Tehran Publication

[8] Bahreini, Seyyed Hassan & Tajibakhsh, Golnar, "Definition of Territory in Urban Design", Honar-Haye-Ziba Journal, University of Tehran Publication, Vol.6

[9] Kar, Alan, "Positive Psychology of Happiness Science and Human Resources", Sharifi, Hassan Pasha & Najafi Zand, Jaafar

[10] Simpson, John, "A space for playing", RIBA publication, 1997, Second Edition

[11] Lawson, Bryan," The Language of Space", London , Butterworth-Heinemann.

[12] Habermas, J., " The structural transformation of public sphere: An inquiry into a category of bourgeois society.", Cambridge: Polity Press. 16.Madanipour, A. (2003). Public