SESSION

INFORMATION AND KNOWLEDGE ENGINEERING AND MANAGEMENT + FEATURE EXTRACTION AND AI

Chair(s)

TBA

Modeling Shared Drive Utilization Using Stochastic Techniques

Margret T. Martin, Sarah G. Nurre and Michael R. Grimaila, *Senior Member*, *IEEE*

***Abstract*—Information Technology (IT) units provide electronic file shared drives for utilization by personnel in their organization. This shared electronic storage space is used for a wide variety of reasons (e.g., archival, collaboration, backups, dissemination) and is generally focused on providing areas for collaboration, as well as to augment the primary storage disk space located within each user's computer system. The ways in which shared drives are utilized are highly dependent upon the organizational mission, who can access shared resources, the stability of the user population, end user roles, and the data retention policies enforced by the IT unit. The goal of this research is to understand what happens to information in shared disk storage within an academic institution as a function of time. Academic organizations are unique due to the transitory nature of the user population (e.g., students arrive and depart each year) and by the various roles that exist within the school. By examining the information lifecycle, we can gain insight into the differing perspectives between end users and IT units, the validity of assumptions about information rot and data aging, and develop an understanding how shared storage space is managed.**

In this paper, we evaluate the utilization of a file-share server used to manage official records within an academic organization and use Discrete Markov Chains to model and simulate the movement of stored data over time as a function of policy within an academic organization. The results show that different IT policies have a dramatic impact on the accumulation of information contained within shared storage space and that organizations should incorporate both the perspectives of the end users and the IT unit when developing organizational policies regarding the use of shared storage space.

***Index Terms*—Information Archival; Information Aging; Data Rot; Information Management; Records Management**

I. Introduction

Virtually all modern organizations have embedded Information and Communication Technologies (ICTs) into their core mission processes as a means to increase operational efficiency, improve decision quality, and reduce operational costs. Within an organization, the Information Technology (IT) unit is responsible for managing and maintaining the organizational ICT resources (e.g., computers, servers, and voice and data networks).

Academic organizations often leverage ICTs in support of the delivery of education to both in-residence and distance learning students, in support research activities, and to enable administration to be conducted in a cost effective, efficient manner. The daily activities of the administration, faculty, staff, and students of the modern academic institution require reliable and usable ICTs to properly attain their mission.

As such, the policies enforced by IT units when managing ICT resources within an academic organization are critical to mission success. One of the primary functions of the IT unit is the management of file-sharing servers, which enable the academic mission and facilitate collaboration within, across, and between institutional units and external collaborators.

End users within organizations have an insatiable demand for storage space. When new shared disk space is provided, inevitably it fills to capacity within a short period of time requiring management of quotas to be enforced so that a small number of users do not monopolize the shared resource. However, enforcing quotas on all users can interfere with the education and research mission of an academic organization when a user has a justifiable need for a large amount of temporary disk space and create a management burden to temporarily allow the user the required space. There are always tradeoffs between organizational policy, perceived user satisfaction, and management costs.

In this paper, we seek to answer the question "What happens to information stored on these shared drives as it ages from year-to-year?" within an academic organization. To answer this question, we first examine the utilization of shared disk storage space and then develop a discrete-time Markov Chain model used to simulate the evolution of stored information. The remainder of this paper is presented as follows. After a brief description and analysis of a shared storage drive within an academic organization under study and development of a conceptual model in Section II, we present the background of using Markov Chain models in Section III. In Section IV, we introduce a model that is used to simulate the impact that different information retention policies have on information storage over a five year period. Section V presents an analysis of the results, and is followed by concluding remarks in Section VI.

II. Conceptual Model Development

In order to accomplish the research objectives, we needed to develop a conceptual model of a shared storage drive. The first step in this model development was to examine the actual disk utilization of shared storage space used within an academic organization. We chose to examine only one of the many shared storage drives contained within the academic organization. The "official records" drive was chosen for analysis because of it clear purpose (i.e., it serves as a repository for the organizations official records) and the limited number of authorized users who have access to the drive. Table 1 below shows an overview of the official records

drive and Table 2 below shows the summary statistics for this drive. Metadata describing the drive was collected by the IT unit using a Microsoft power shell script and subsequently inserted into Microsoft Access database to facilitate analysis.

Table 1. Official Records Drive Contents Overview

Total Number of Directories and Subdirectories	Total Number of Files	Total Space (Bytes)	Number of Unique File Extensions
96	49,885	44,530,209,647	365

Table 2. Summary Statistics for Official Records Drive

Mean	Standard Deviation	Minimum Size (Bytes)	Maximum Size (Bytes)
892657.30	5076508.94	2	334370852

There are several assumptions written below regarding the analysis of this data within the official records drive:

- Electronic files are counted based upon the number of unique extensions.
- Files are arranged based upon the calendar year the file was created. Even if a file was created in 2012, but placed on the shared drive/repository, in 2014, it counts as part of the 2012 data.
- File size is summed by year to determine the amount of storage space utilized.
- Data collection ended in May 2014, so the 2014 data point does not contain an entire year of data.
- It is important to realize not all data from each year is captured. This is because the IT unit clean-up policies subject each drive to certain deletion rates, based on when students graduate, preparations for inspections, and/or limitations pertaining to community drive space.
- The data is analyzed based upon the types of files users created, respective creation years, and is displayed using pie charts. This was accomplished by grouping files by category, which incorporated multiple extension types. These graphs provide a unique perspective of what was kept for archival purposes versus what was being created by the institution.

It is also important to note limitations of the analysis. Although the organizations maintain data for multiple years, the resources required to retrieve this data were not available. As a consequence, the data presented is a snap-shot of each drive's contents from May 2014. Ultimately, this was a single-point-in-time analysis. A request was made to the IT unit of the organization to provide the creator of each file; however the administrative support required to collect this data was not available. As a result, it was difficult to determine how many different users were contributing to the shared drive.

Our main intent in collecting data from the official records drive was to ask, "What is the organization working on and how are records preserving this as the spirit of the mission or as transparency in operations?" In order to understand where specific disconnects exists, we examined the data from a few different perspectives. We first sought to compare the growth of records in each of these file-sharing repositories to literature findings which state information grows at an exponential rate. We were interested to see if the collected data would support these findings. We hypothesized that while this may be true for individual user disks, we did not believe this would be true for a shared disk whose purpose was to house organizational official records. We suspected that population of the official records shared drive space would be driven by rate at which official records are created within the organization. Figure 1 below shows the cumulative number of bytes stored in the official records repository by year from 2005-2014.

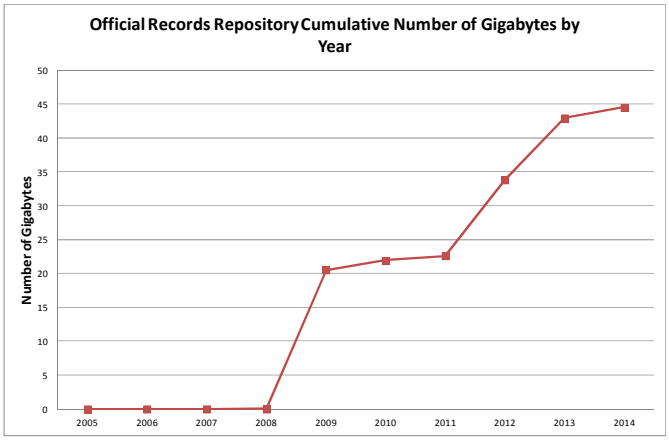


Figure 1. Official Records Repository (Drive) Cumulative Number of Gigabytes as a Function of Year

Notice that there are more dramatic increases in cumulative data in the period from 2008-2009 and 2011-2013. After consulting with Subject Matter Experts (SMEs), we determined that this can be explained by the fact that more data is added to the official records repository just prior to and during inspection years (2009 and 2012). Thus our belief that official records were added to the drive as they were created was incorrect. Instead, it appears that just prior to records management compliance inspections that the organization "rushes" to add files to the official records drive so that they will pass inspections. This analysis explains the organizational behavior with regard to the use of the official records shared drive. This is evident by the way the IT unit manages available storage space, the number and types of users, and the purpose of the drive. While it's possible to fit a trend line to the data, it would not add useful insight into understanding the growth of records in the official records repository because of its inherently piece-linear nature. Looking at the data from another lens, Figure 2 shows the files on the official archives drive by creation date. Notice the large number of files added in the years 2006, 2009, and 2012. This confirms the earlier observation that large numbers of files are added just before and during inspection years. Archiving and records management tend to be passive activities and the right resources and people must be available to conduct these efforts or else they will not occur until there is a tangible penalty (e.g., a failed inspection) levied against the organization.

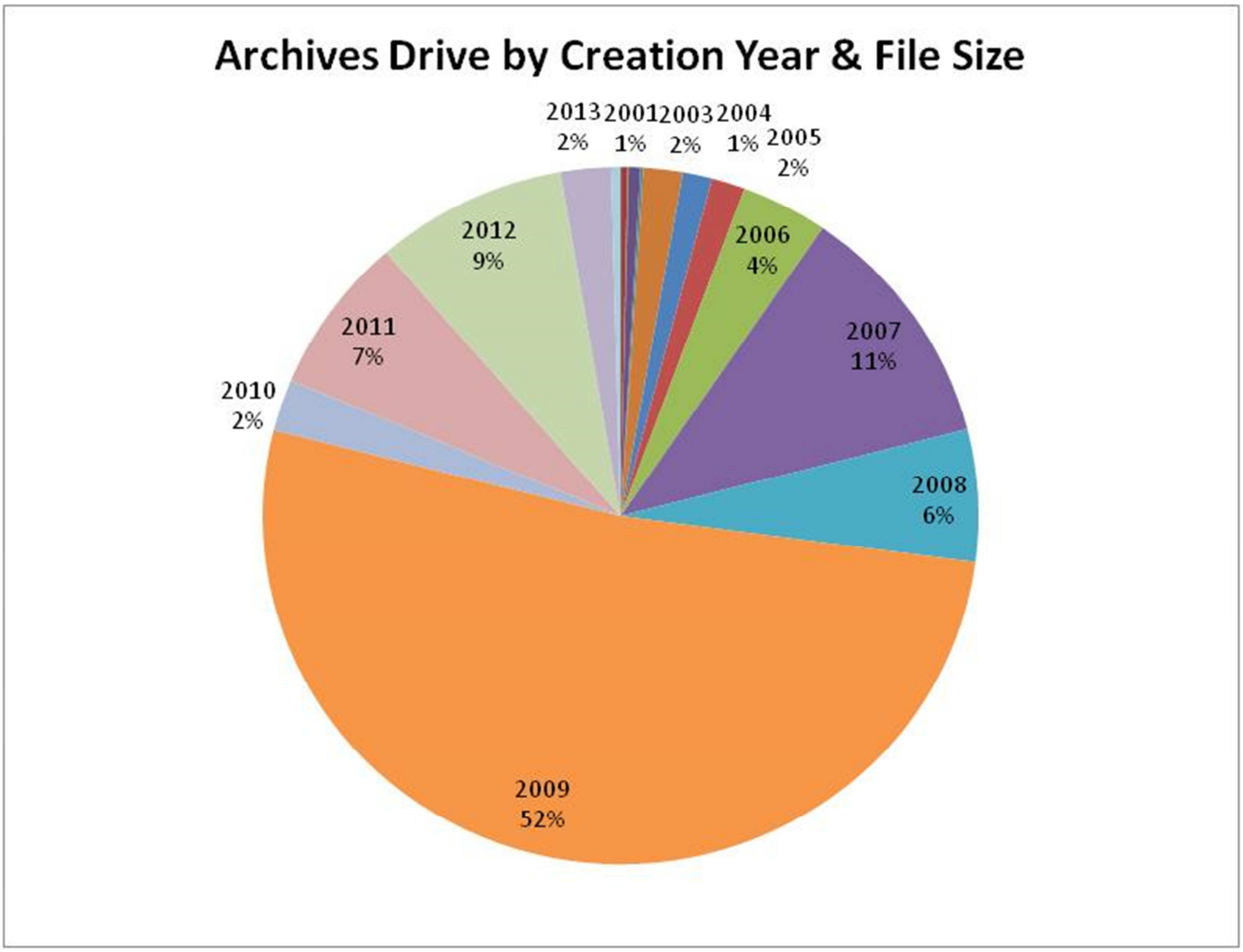


Figure 2. Official Records Repository (Drive) Files by Creation Year and File Size

We now examine the type of data stored on the official records drive. In this case, the number of files and their extensions were grouped by purpose and technological medium. For example, the "image" category encompasses file extensions such as .tif, .png, ,fig, .gif, .bmp, and .jpg. Note that only the top 25 extension types were considered for grouping purposes and the remainder was grouped into an "other" category. In the case of the official records repository, shown in Figure 3, the largest groupings consisted of 45% Adobe PDF files, 15% Microsoft Word documents, 7% computer languages, and 7% PowerPoint presentations. Only 15% were captured in the "Other" category. This was as expected, as the official records drive typically contains institutional records which are typically archived in the above common formats as opposed to general purpose drives that may contain a much broader spectrum of file types.

The analysis of data by extension types provided unique insight into what is important to the organization as a whole. A more in-depth analysis would have included grouping the information by user role, but unfortunately, the required data was not available.

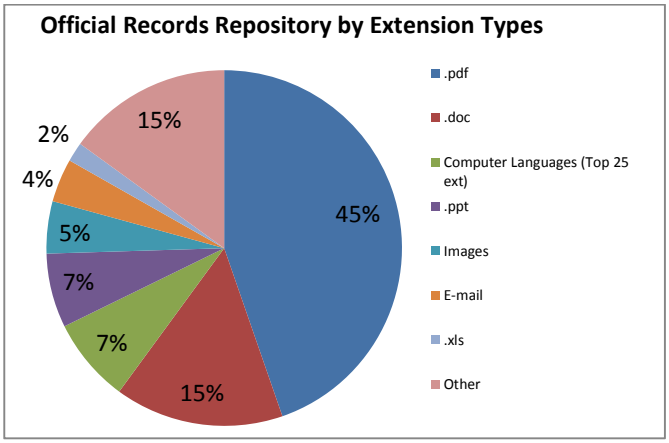


Figure 3. Official Records Repository (Drive) by Extension Type

IT governance principles explain how data is managed based upon who created it and what their role is within the organization hierarchy. For example, leadership files are seldom purged or deleted by the IT unit. In contrast, files created by students are regularly viewed for deletion.

The next step was to develop a generalized model to demonstrate an overview of information flow on the shared drive as shown in Figure 4. A time series analysis was conducted from the initial state of the drive in 2009, through the years 2010 through 2014. The initial input for this model system was labeled "Initial Info," and represents all of the electronic data including Word documents, Excel files, PowerPoint presentations that were created, posted, and last modified in, or prior to, the year 2009.

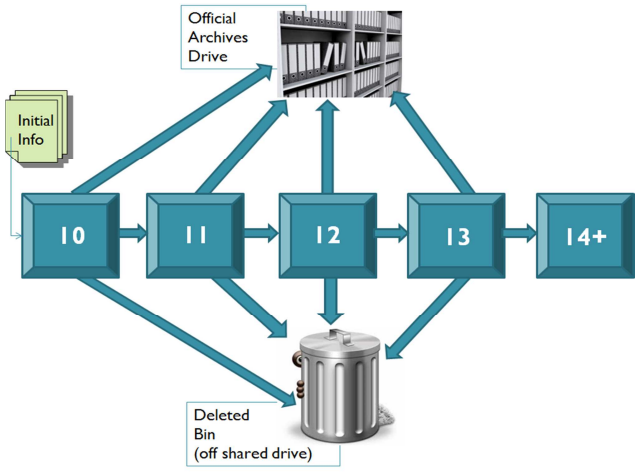


Figure 4. Generalized View of File-Sharing Server System

The model shown in Figure 4 is simplified and is used to evaluate what happens to "old information" as time increments from one calendar year to the next, given the possible paths this information can travel for the year range from 2010-2014. It is important to note that in this simplified model no additional information is added to the system in 2010 or thereafter. End users and IT Personnel in the academic user community have the following 3 choices:

1. Allow the initial data to stay on the shared drive (either for its inherent value, because a shared drive "clean-up" has not be conducted, or because users left it there). The option to keep "old information" is illustrated in Figure 4 by the horizontal lines running from one calendar year node to the next.
2. Archive the 2009 information onto the shared official records drive. Because the official records drive is not visible to a great majority of users, this is accomplished via the assistance of the academic institution's IT directorate and is depicted in the model by the arrows from each calendar year to the official records drive.
3. Purge the initial information. This option is demonstrated by the lines pointing to the trash receptacle from each calendar year node, which means it has moved off of the shared server space. Users, of course, are free to back-up their own information on personal storage devices as they choose and as allowed by the institution.

In a real world system, many issues stem from the fact that users add documents, data, and information to the system each year, but this difference is what allows this model to serve as a simple replication. Now that the basic, feasible paths are explained, the next step is to apply stochastic modeling concepts.

III. Markov Chain Models

Previous research has been conducted using Markov Chains in order to evaluate ICT phenomena. For example, Yossef et al. [1] used a one-dimensional Markov Process, or random walk, to approximate certain aggregate queries, such as search engine usage and the proportion of pages belonging to .com or other domains. Attempts to estimate the size of a domain, or estimating the fraction of web pages covered by a search engine are both efficient and require very limited resources. Thain et al. explained that end users and systems administrators have "two distinct roles to play" and the importance of IT professionals being able to apply set constraints while users must be given elements of freedom to work as their mission requires without extreme limitations or constraining policy requirements [2]. This can involve the implementation of distributed storage systems with two distinct intents, or services: storing data vs. organizing directories. Ultimately, this research highlights the idea that administrators shouldn't care about the purpose for why a user is employing a file server, with the exception of security reasons and resource policies. Flexible policies should be set in place to lead to new modes of interactions for users.

The distinction between the interpreted value (or usefulness of the data) vs. an IT administrator's due diligence in managing limited storage, server space can lead to some interesting assumptions from both ends of the spectrum. Whether it's accidental or intentional deletion of data, it's important to realize that any risks and faults, albeit latent or visible, are memoryless according to Baker et al, and similar to a Markov Chain [3]. Additionally, two "dangerous assumptions" that the article mentions are an unlimited budget assumption and human error which are disconnects when conduction long-term digital preservation. Work is this realm is in high-demand. This is evident by works such as Fessant, et al who specifically recommend further analysis of peer-to-peer networks [4] and Z. Ge et al [5], as well as research pertaining to cost-effective file migrations of servers [6].

IV. A Discrete Markov Chain Model

To stochastically model this system, we developed a discrete-time Markov Chain. A Markov Chain is a mathematical system that undergoes transitions from one state to another and is deemed 'memoryless' such that what happens in the future depends solely upon the current state and the probabilistic determination of the projected path.

Note that in Figure 5, the model was updated so that the official records drive (the O:/ drive) now appears to 'recycle' back onto the calendar year nodes. Modifying the archival option for the data to a transient state at each calendar year

node is important because the official archives drive does count towards the institution's storage limit authorized, in terms of shared server space and it shows that information which was once archived can be moved back onto the drive so it will continue to the next calendar year or it is moved to the deletion bin. This is an important caveat pertaining to the institution's records management regulations. The only two absorbing states in this Markov Chain are the 2014+ node and the deleted items bin. Once a transition into an absorbing state occurs, the information, or data, will remain there forever.

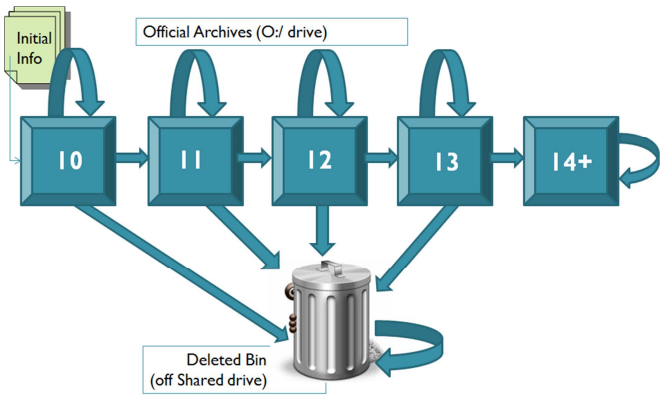


Figure 5. Markov Chain Model of Information Flow

A. Assumptions

Because longitudinal data was unavailable for the official records drive, a table of important perspectives and assumptions was devised which led to an analysis stemming from two different lenses by which to view the shared drives as shown in Table 3 below.

Table 3. Information Retention Policy Perspectives/Assumptions

Perspectives	IT Perspective	Only so much storage space is available in the institution; IT's job is to archive official records & manage shared drive space (but not to determine the value of the information).
	Users Perspective	There is value added by keeping information (i.e. mission requirements, "just in case we need it"); Generally purge information only when prompted to do so.
Assumptions	General	A large percentage of information "rots" with time (especially when it is not updated or utilized).
	Even-numbered Years	Records Management inspections usually occur during even-numbered years; IT is more apt to archive e-files on the official archives drive then.
	Odd-numbered Years	Server space fills up on these years, so IT initiates 'clean-sweeps' to encourage users to purge unnecessary information.

B. Analysis

After the academic institution's IT and user perspectives were realized, two Markov Chain transition matrices [7] using randomized probabilities were qualitatively generated using SME inputs. These values had to be assumed due to the lack of availability of the time series data required to properly estimate these parameters. Table 4 shows the probability transition matrix representing the IT personnel perspective and Table 5 shows the probability transition matrix representing the academic user perspective. The models were verified by multiple sources within the IT realm, including a PhD with a background in Information Sciences.

As an example, consider Table 4 which shows the P-matrix from the IT personnel perspective. In this case, the percentage of 2009 information that is likely to stay on the server from 2010 to 2011 is 86%, while 93% of this data would be retained if users exercised their organizational behavior. The transition matrices provide estimated probabilities associated with the transition from one state, or calendar year (the headings on the left of the matrix) to another (the headings along the top of the matrix).

In both matrices, the far right hand column shows the likelihood of information being deleted off the file-shared drive. The IT unit's philosophy is based on the idea a specified amount of storage space exists, so this directorate must actively delete electronic records after a certain time period assuming the concept of information rot. Users do not abide by this same rationale, and instead, only exert clean-up efforts when prompted to do so, by IT personnel working with the institution's leadership.

Table 4. P-Matrix for IT Personnel Perspective

\To From\	I0	I1	I2	I3	I4	Delete
I0	0.04	0.86	0	0	0	0.1
I1	0	0.02	0.7	0	0	0.28
I2	0	0	0.03	0.55	0	0.42
I3	0	0	0	0.02	0.43	0.55
I4	0	0	0	0	1	0
Delete	0	0	0	0	0	1

Table 5. P-Matrix for Academic User Perspective

\To From\	I0	I1	I2	I3	I4	Delete
I0	0.02	0.93	0	0	0	0.05
I1	0	0.02	0.9	0	0	0.08
I2	0	0	0.01	0.94	0	0.05
I3	0	0	0	0.01	0.91	0.08
I4	0	0	0	0	1	0
Delete	0	0	0	0	0	1

The numbers along the primary diagonal of each matrix show the probability of archival, which are slightly elevated during records management inspections years, especially because the IT unit owns and is responsible for the records management program. The remaining positive values, all less than 1.0, show the amount of 2009 information that moves from one calendar year to the next on the shared drive. Notice by the lack of participation in cleaning and maintaining the drives and information, that users have a much higher probability of letting 2009 information stay on the drive as opposed to removing it. Using occupancy probability matrix equations [8], the aforementioned transition matrices (i.e. P-

matrices) were evaluated to calculate the π values and n-step transition matrices associated with the number of time-steps to move all the initial information from 2009 into the absorbing states. Figure 6 shows the simulated information retention rates as a function of time for years 2010-2013 (corresponding to n=2, 3, 4, and 5) for IT operations personnel when using the P-matrix shown in Table 4. Figure 7 shows the simulated information retention rates as a function of time for years 2010-2013 for academic users when using the P-matrix shown in Table 5.

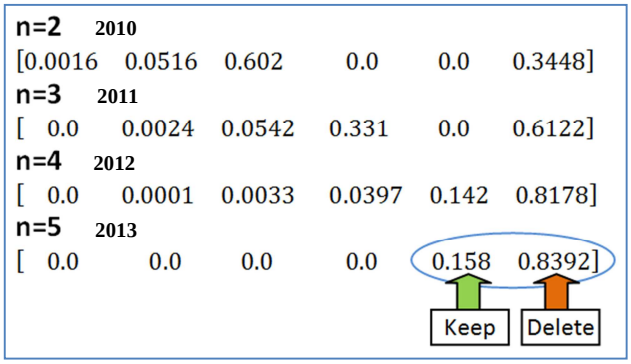


Figure 6. Information Retention Rates for IT Operations Personnel

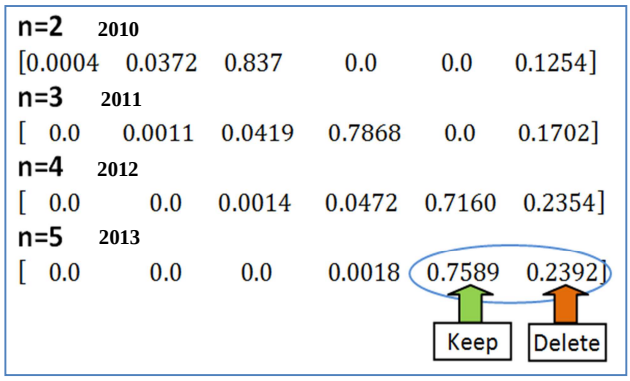


Figure 7. Information Retention Rates for Academic Users

Evaluating time steps, from 2010 through 2013, simulates the movement of information on the shared drive from year to year and reveals the information retention after 4 years (at n=5) for the official archives. Assuming the IT and user 'policies' ran their courses separate from one other, this analysis demonstrates that after 6 years, the IT Directorate would delete approximately 83.92% of the 2009 data off the shared drive and 15.8% would still remain on the drive. Relying on a "users decide" policy, 75.89% of the initial 2009 information would still remain on the shared drive while only 23.92% would be deleted.

V. Conclusions

The intent of this research was to better understand what happens to information on a shared drive as it ages and study the amount of information that accumulates by evaluating differing policies, or perspectives. In order to scope this project appropriately and remain sensitive to the workload required by the institution's IT unit, interviews were conducted with the IT Director and various SMEs. A discrete-time Markov Chain was created and an analysis was conducting to determine what the anticipated information flow looks like from year-to-year.

The analysis behind this discrete-time Markov Chain and associated probability matrices demonstrates a genuine disconnect in the way users and IT personnel view and treat shared server space based on the keep-to-delete ratios. When using the IT operations personnel policies, the keep-to-delete ratio was approximately 16:84. In contrast, when using the academic user's policies we obtain a keep-to-delete ratio of 76:24 over the same time period. This is not surprising and notionally matches the behavior observed in the organization. Users will store files perpetually, if allowed, so that they will have another backup of their critical records.

While both perspectives have valid viewpoints, the analysis behind this discrete-time Markov Chain and associated probability matrices demonstrates a genuine disconnect in the way users and IT personnel view and treat shared server space based on the keep-to-delete ratios. Because so much untouched, unreferenced, and unmodified information has a way of accumulating on these servers, it is no wonder there is seldom enough storage space available at this academic institution and that the shared drive icons turn red so frequently, signaling they are close to maximum capacity as seen in the drive utilization shown in Fig. 8.

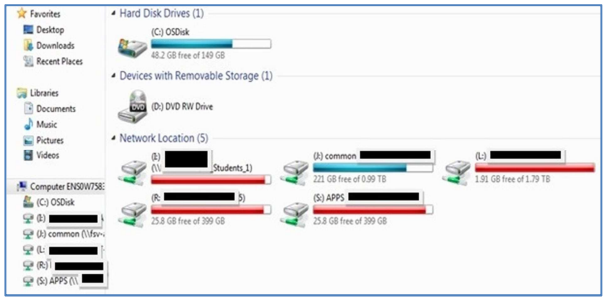


Figure 8. Academic Shared Server Space Utilization by Drive

It is recommended that additional operations research concepts be applied to this research, such as Markov Decision Processes and Bonus-Malus systems which can allow differing costs and incentives to be associated with the decisions to better understand organizational behavior, trends, and related policies. As previously discussed, even if the IT policy determines that 16% of old data from 6 years prior should remain on the storage drive, and assuming that same amount of information is retained year after year, nearly 20% of the server will be comprised of data 5+ years old. Naturally, this is just hypothetical, but would not be feasible if the amount of storage space ceases to increase.

Many of the assumptions in this paper derive from the mathematical probabilities related to human behavior and the older information and data becomes, the higher probability it has to become archived or deleted based on obsolescence. In addition, it's important for follow-up research to be conducted to truly evaluate which files are used, accessed, and modified

vs. that which is retained and never retrieved, which may include many of the archives. An important follow-on question is: Can 'carrying costs' be associated with the movement of information on a file-sharing drive in order to create a more effective policy?

VI. Disclaimer

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the U.S. Government.

References

[1] Z.B. Yoseff, A. Berg, S. Chien, J. Fakcharoenphol, D. Weitz. (2000.) Approximating Aggregate Queries about Web Pages via Random Walks. Proceedings of the 26th VLDB Conference, Cairo, Egypt. pp 535-544. [Online]. Available: http://www.vldb.org/conf/2000/P535.pdf

[2] D. Thain, S. Klous, J. Wozniak, P. Brenner, A. Striegel, J. Izaguierre. (2005, Nov.) Separating Abstractions from Resources in a Tactical Storage System. Seattle, Washington. [Online]. Available: https://www3.nd.edu/~ccl/research/papers/tss-sc05.pdf

[3] M. Baker, M. Shah, D. Rosenthal, M. Roussopoulos, P. Maniatis, T. Giuli, and P. Bungale. (2005, Aug.) A Fresh Look at the Reliability of Long-Term Digital Storage. ACM SIGOPS Operating Systems Review. Vol. 40. No. 4. pp 1-14. [Online]. Available: http://arxiv.org/pdf/cs/0508130.pdf

[4] F.L. Fessant, S. Handurukande, A.-M. Kermarrec, L. Massoulie. (2004.) Clustering in Peer-to-Peer File Sharing Workloads. IPTPS pp. 217-226

[5] Z. Ge, D.R. Figueiredo, S. Jaiswal, J. Kurose, D. Towsley. (2003). Modeling Peer-Peer File Sharing Systems. *Proceedings of IEEE INFOCOM*. pp. 2188-2198.

[6] B. Gavish and O.R. Liu Sheng. (1990, Feb.) Dynamic File Migration in Distributed Computer Systems. Management of Computing, Vol. 33, No. 2, pp. 177-189.

[7] Sheldon M. Ross. *Introduction to Probability Models, 9th edition.* Academic Press, 2006.

[8] C. Richard Cassady, J.A. Nachlas. Probability Models in Operations Research. (2008, Aug.) CRC Press.

Collaborative Shared Awareness: Human-AI Collaboration

James A. Crowder, John N. Carbone
Raytheon Intelligence, Information, and Services
16800 E. Centretech Parkway, Aurora, Colorado 80011

Abstract - *The ability to reason within an autonomous information processing system denotes the ability to infer about information, knowledge, observations and experiences, and affect changes within the system to perform new tasks previously unknown, or to perform tasks already learned more efficiently and effectively [Crowder 1996]. The act of reasoning and inferring allows an autonomous system to construct or modify representations of concepts or knowledge that the system is experiencing and learning. Reasoning allows an Artificially Intelligent System (AIS) to fill in skeletal or incomplete information or specifications about one or more of its domains (self-assessment). The research described here details architectures and algorithms for a cognitive system of Intelligent information Software Agents (ISAs) to facilitate Collaborative Shared Situational Awareness between humans and systems [Crowder, Scally, and Bonato 2012]. The original purpose of this research is to design the algorithms and architectures needed for a system of heterogeneous software agents to autonomously mimic human reasoning in the cognitive ways brain processes information and develops knowledge [Crowder 2010][Carbone 2010], while simultaneously providing the human operators with the ability to monitor the autonomous system and allowing the operator to provide feedback and instruction to the system to facilitate improvement (human-AI collaboration). This knowledge takes the form of answering questions and explaining situations that the autonomous system might encounter. ISAs are persistent software components, called Cognitrons (Cognitive Perceptrons), which perceive, reason, act, and communicate [Crowder, Carbone, Friess 2013]. The research described here entails the design and implementation of the ISA Cognitron algorithms and the architecture required to provide a system capable of autonomously managing a complex network of assets to enhance Situational Awareness and optimize network asset utilization. This research includes algorithms and architectures that facilitates system learning from interaction with operators via the Human Mentored Software. This system is called the Cognitive, Interactive Training Environment (CITE) and will allows Human Interaction Learning [Crowder and Friess 2012].*

Keywords: Knowledge Essence, Artificial Reliability, Self-Evolvable Systems, Knowledge Relativity Threads, Physical Mechanics, Metacognition.

1. Introduction

As global populations and societies continually reach epic proportions and resources become continuously constrained, Human Needs Engineering (HUMANE) is not only required but needs to be an inherent component to all engineering disciplines. Additionally, as proportions rapidly increase and resources rapidly decline, more effective, real-time, automated, and dynamically human interactive systems become required. Recent research within highly automated and autonomous domains shows promise mitigating the need for critical intelligent infrastructure to improve human-system collaboration, awareness, and quality of service (QOS). Hence, to improve decision making, an Artificially Intelligent System (AIS), in order to be truly autonomous, is provided with a real-time, human like, cognition-based framework for information discovery, decomposition, reduction, normalization, encoding, and memory recall (knowledge assimilation and construction)[Carbone 2010]. To achieve efficient human-system knowledge/needs collaboration, these currently researched cognitive systems work to integrate information into their Cognitive Conceptual Ontology (Crowder, Taylor, and Raskin 2012) in order to be able to "think" about, correlate and integrate information content into internal memories. When describing how science integrates with information theory, Brillouin [Brillouin 2004] defined knowledge succinctly as resulting from a certain amount of thinking, distinct from information content which initially had no value, was the "result of choice," and consisted of simply raw material, a mere collection of data. Brillouin concluded that a hundred random sentences from a newspaper, or a line of Shakespeare, or even a theorem of Einstein have exactly the same information value and had "no value" until effort of thought was applied to turn information content into knowledge. In the Health industry decision-making is a great concern due to the information content ambiguity and ramifications of inferences made erroneously. Often there can be serious consequences

when actions are taken based upon subsequent incorrect recommendations. Decision-making can be influenced prior to inaccurate inferences being detected and/or even corrected. Hence, underlying the data fusion domain is the challenge of creating actionable knowledge from information content harnessed from an environment of vast, exponentially growing structured and unstructured sources of rich complex interrelated cross-domain data[Llinas, Bowman et al. 2004]. This is a major challenge for AI systems that must deal with ambiguity in human-based collaboration and operator-based assistance. Therefore, in this paper we discuss engineering architecture and concepts of human-artificially cognitive systems and a joint collaboration environment that could allow human mentors to develop cognitive trust and reliance of collaborative AI system systems within a populace. These systems would be providing humans timely and reliable knowledge rapidly mitigating their daily needs, allowing each to not only learn from each other, but operate in modes that utilize the strengths of both. This includes cognitive procedural memory development that will allow improvement in attitudes and knowledge about the value artificial life forms and autonomous systems.

2. The Essence of Meaning

Intelligence reveals itself in a variety of ways, including the ability to adapt to unknown situations or changing environments. Without the ability to adapt to new situations, an intelligent system is left to rely on a previously-written set of rules, making collaboration difficult, since the AI System (AIS) cannot keep up with the human operator who has the ability to adapt to new situations. If we truly desire to design and implement collaborative AI Systems (AIS), they cannot require precisely-defined sets of rules for every possible contingency. The questions then become:

- ***How does an AI system construct good representations for tasks and knowledge as it is in the process of learning the task or knowledge?***
- ***What are the characteristics of a good representation of a new task or a new piece of knowledge?***
- ***How do these characteristics and the need to adapt to entirely new situations and knowledge affect the learning process?***

Given the AI system has bounded resources, it would need to react utilizing the concepts of Cognitive Economy to create a Bounded Rationality set of goals to solve a particular problem or situation. These are:

1. The size of the feature set – how many "features" are required to define the success of each task
2. The "fuzzy" relevance of each feature for the tasks
3. The preservation of necessary distinctions for success in each task

The AIS's cognitive components would autonomously define, for each ISA, a Banach Space for that ISA's goals and tasks and would then consider the set of ISA Banach Spaces as a set of bounded variations, the sequence of which (through ISA collaboration) produces an acceptable solution to the situation(s) or task(s) at hand.

The Cognitive Economy methods will be described and a discussion will be provided, illustrating how these Cognitive Economy and Bounded Rationality concepts affect the overall learning aspects of an autonomous AIS.

In addition, when considering autonomous AIS, we must consider its need to interact and learn from its environment, and we have to ask ourselves "what is reality?" We have to establish how the AIS would interpret their reality. One of the issues that humans deal with that assists in their understanding of reality, or their world around them and how they need to interact, is their concept of "Locus of Control." **Locus of control** is a term in psychology that refers to a person's belief about what causes the events in their life, either in general or in specific areas such as health or academics. Understanding of the concept was developed by Rotter [Rotter 1954], and has since become an important aspect of personality studies.

2.1 AIS Constructivist Learning

Constructive psychology is a meta-theory that integrates different schools of thought. According to the above cited article:

Hans Vaihinger (1852-1933) asserted that people develop "workable fictions". This is his philosophy of

"As if" such as mathematical infinity or God. Alfred Korzybski's (1879-1950) "System of Semantics" focused on the role of the speaker in assigning meaning to events. Thus, constructivists thought that human beings operated on the basis of symbolic or linguistic constructs that help navigate the world without contacting it in any simple or direct way. Postmodern thinkers assert that constructions are viable to the extent that they help us live our lives meaningfully and find validation in shared understandings of others. We live in a world constituted by multiple realities social realities, no one of which can claim to be "objectively" true across persons, cultures, or historical epochs. Instead, the constructions on the basis of which we live are at best provisional ways of organizing our "selves" and our activities, which could under other circumstances, be constituted quite differently.

For AIS with Constructivist Learning, the AIS cognitive learning process would be a building (or construction) process in which the AIS cognitive system builds an internal illustration of its learned knowledge-base, based on its experiences and personal interpretation (fuzzy inferences and conceptual ontology [Raskin & Taylor 2010a and Taylor & Raskin 2011a] of its experiences. AIS Knowledge Representation and Knowledge Relativity Threads [Carbone 2010][Crowder and Carbone, 2011c], within AIS cognitive system memories would be continually open to modification, and the structures and linkages formed within AIS short-term, long-term, and emotional memories [Crowder and Friess, 2010b], along with its Knowledge Relativity Threads [Crowder and Carbone 2011c], would then form the bases for which knowledge structures would be created and attached to AIS memories.

One of the results of the Constructivist Learning process with the AIS would be to gradually change its "Locus of Control" for a given situation or topic, from external (the system needing external input to make sense, or infer, about its environment) to internal (the AIS having the cumulative constructive knowledge-based of information, knowledge, context, and inferences to handle a given situation internally); meaning the AIS is able to make relevant and meaningful decisions and inferences about a situation or topic without outside knowledge or involvement. This becomes extremely important for completely autonomous AIS.

2.2 Bounded Conceptual Rationality (Cognitive Economy)

Bounded rationality is a concept within cognitive science that deals with decision-making in humans [LaBar and Capeza 2006]. Bounded rationality is the notion that individuals are limited by the information they have available (both internally and externally), the finite amount of time they have in any situation, and the cognitive limitations of their own skills. Given these limitations, decision making becomes an exercise in finding an optimal choice given the information available. Because there is not infinite information, infinite time, nor infinite cognitive skills, humans apply their rationality after simplifying the choices available, i.e., they bound the problem to be solved into the simplest cognitive choices possible [Jones 1999].

Any AIS must suffer the same issues. An autonomous system, by definition, has limited cognitive skills, limited memory, and limited access to information. The Locus of Control concepts discussed earlier assist AIS in determining which situations can be handled internally vs. externally, but still in any situation there is limited information, time, and cognitive abilities. This is particularly true if the system is dealing with multiple situations simultaneously. In order for the system to not become overloaded, we believe autonomous systems must employ strategies similar to human bounded rationality in order to deal with unknown and multiple situations they find themselves in. This involves creating mathematical constructs that can be utilized to mimic the notion of bounded rationality within autonomous AIS.

For this we look to Banach Space theory, tied into Constructivist Learning concepts [Botella 2011] for autonomous AIS. As concepts are learned and stored in the AIS conceptual ontology [Raskin & Taylor 2010a], Banach Spaces are defined that are used to bound the rationality choices or domains for that concept. As we "construct" these concepts and the Banach Spaces that bound them, the combination of Banach Spaces then defines the Conceptual Rationality for the Autonomous AIS. Figure 1 illustrates this concept. These Banach Spaces that define the bounds for each learned concept are utilized when the AIS must reason, or perform decision making. When there are restricting limitations on time, resources (as determined by the

resource manager, e.g., artificial prefrontal cortex), and available information, the bounds of these Banach Spaces would be tightened or loosened to allow the AIS to deal with multiple situations, or situations that are time critical. This allows AIS to decide what is a "good enough" solution to a given problem or set of problems, and to adjudicate between competing resources, priorities and overall goals

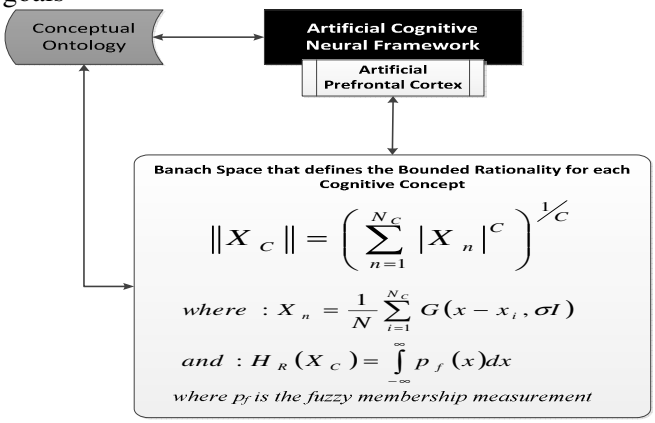


Figure 1 – AIS Bounded Conceptual Rationality

3. Human-AI Collaboration

3.1 Cognitive Architectures for Human-AI Communication

Here we describe an Intelligent information Software Agent (ISA)-based cognitive system that provides a distributed, extensible, and dynamically changing, learning, and self-adapting processing environment. This system, called the Polymorphic, Evolving, Neural Learning and Processing Environment (PENLPE). PENLPE represents a massively parallel, highly interconnected network of loosely coupled, relatively simple processing elements; Intelligent information Software Agents (ISAs), called "experts," in a hybrid fuzzy, genetic neural system of "M" experts architecture [Crowder 2010a]. The purpose of PENLPE is to provide a hybrid neural processing environment that is adaptable to a variety of classes of applications (e.g., language processing, signal detection, sensor fusion, inductive and deductive inference, robotics, diagnosis, etc.). The PENLPE architecture is based on a "mixture of experts" methodology. The difference here is that in our architecture, an expert is defined as a particular fuzzy, genetic perceptron ISA object which has been created for a particular algorithm or problem, and thus is an expert at processing a particular type of data in a particular manner. The algorithm(s) for which the perceptron ISA is generated may be predetermined or may have been evolved by the neural system itself. The PENLPE cognitive architecture (Figure 2) takes input from a heterogeneous set of information sources (sensors), facilitates the fusion of the information from these sources, and automatically provides situational assessments. This provides the agent tasking and sub-tasking required for the processing goals and requirements. The impact and benefit of such an autonomous collection system is:

1. Reduction in data acquisition and recognition time.
2. Improved efficiency for autonomous decision support
3. Improved processing and reporting timeliness
4. Improved decision support quality
5. Effective knowledge and decision management.

Designs for the various ISAs and information management algorithms are combined to produce a design capable of providing autonomous cognitive agents, called Cognitive Perceptrons, to automate the situational awareness activities within a robotic system. The main ISA archetypes (see Figure 3) are:

1. The Interface Agent: The Interface Agent assesses the correctness of major decisions and adjusts the decision processes of the Advisor Agents. Interface Agents also accommodate human-in-the-loop structures
2. The Data Steward Agent: This agent acquires raw data from a variety of sources, including sensors, and prepares incoming data for use by other agents. The Data Steward Agent generates and maintains metadata required to find and extract data/information from heterogeneous sources
3. The Reasoner Agent: The Reasoner Agent interacts with the Data Steward and Advisor Agents and utilizes the ontologies and lexicons to automate the development of domain-specific encyclopedias; it provides a mixed source of information and question answering that is used to develop an understanding of questions, answers, and their domains. Reasoner Agents analyze questions and relevant source information to provide answers and to develop cognitive ontology rules for PENLPE.
4. The Analyst Agent: The Analyst Agents are fed by Reasoner Agents and utilize the developed

ontologies and lexicons to expand upon questions and answers learned from collected information.

5. The Advisor Agent: This agent disseminates the right information to the right place at the right time; it provides capabilities that allow collaborative question asking and information sharing by agents and end-users.

Any autonomous information processing and situational awareness agent-based system must consider overall real-time performance issues. It should have the capability to overcome inherent bottlenecks that result from massive volumes of data being generated by the collection sensors or processors transforming the data into information and knowledge [Crowder 2012a].

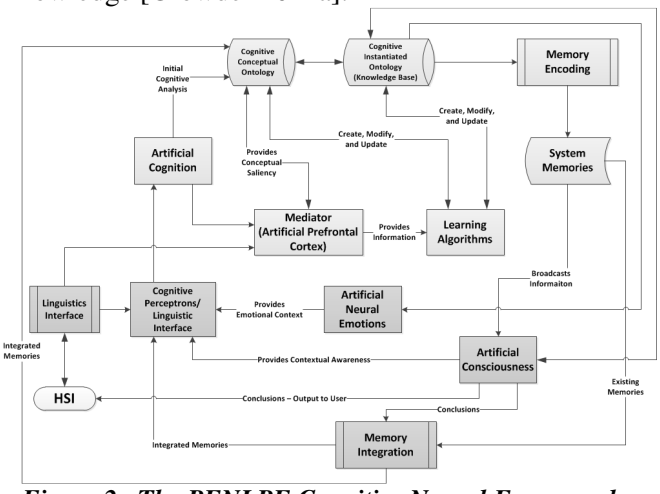


Figure 2– The PENLPE Cognitive Neural Framework

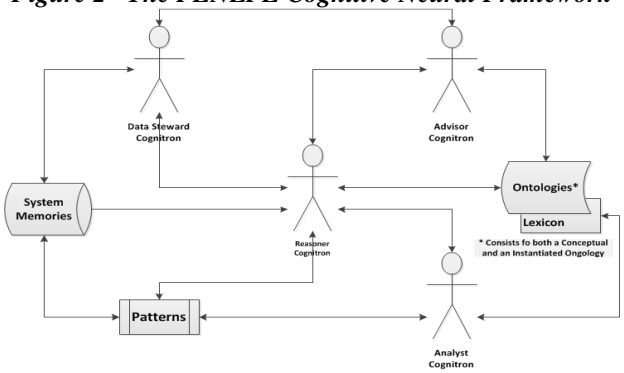


Figure 3 – The ISA Communicative Ecosystem

3.2 Communication for Human-AI Collaboration

Utilizing software to partially or fully automate tasks is now commonplace. However, the capabilities of the software performing these tasks typically do not improve over time (as humans would who were performing the same tasks). We describe here the use of a software system called the Cognitive, Interactive Training Environment (CITE) that learns and improves through the use of a Human Operator acting as a Mentor for the software, until the software is capable of performing the desired operations autonomously and with improvements. CITE provides for Human Interaction Learning (HIL), as the operators role changes from manager to mentor to monitor while the software evolves from learner to performer. One of the purposes of this research is to determine the Levels of Automation of Design and Action and the cognitive software architectures required to allow the system to learn and evolve [Crowder, Carbone, and Friess 2013]. The CITE system (Figure 4) provides effective feedback mechanisms to allow humans to influence PENLPE in a positive way and allows PENLPE's ISAs to learn and improve as they process. The human mentor has the ability to query the system, based on PENLPE's suggestions and then provide feedback as to why a given choice or set of choices was effective or not. PENLPE will provide feedback to the operator to give human mentor an understanding of the process PENLPE utilized to make inferences and decisions. This process of feedback and PENLPE-human mentor interactions provides the operator the insight to develop trust in PENLPE over time and to increase the efficiency of both PENLPE and the human mentor.

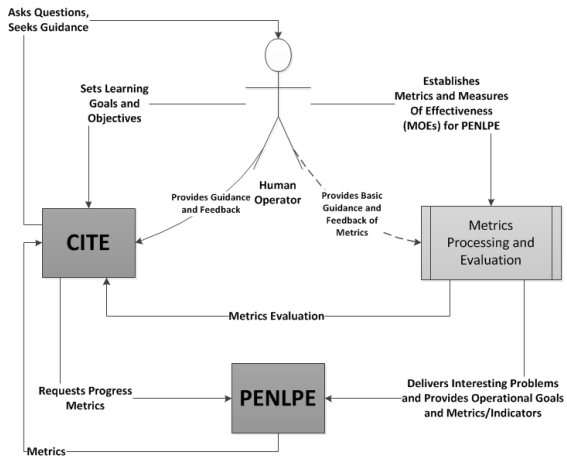


Figure 4 – The CITE Human Mentored Software Environment

4. Conclusions and Discussion

What we have described is a human-artificially cognitive system collaboration environment, CITE, that will allow human mentors to develop cognitive trust and reliant on autonomous system and provide

knowledge products that reflects state of the art cognitive interaction between artificially cognitive systems and humans, providing a new generation of human-machine collaboration, allowing each to not only learn from each other, but operate in modes that utilize the strengths of both. This includes cognitive procedural memory development that will allow improvement in attitudes and knowledge about the value artificial life forms and autonomous systems through cognitive self-awareness, self-evaluation, and self-regulation.

5. References

1. Brillouin, L. 2004. Science and information theory. Dover.
2. Newell, A. 2003. Unified Theories of Cognition. Cambridge MA: Harvard University Press.
3. Crowder, J. A. 1996. Reusable Launch Vehicle Automated Mission Planning Concepts. NASA Report X33/RLV-MP-PHII-1996-005, Lockheed Martin, Littleton, CO.
4. Crowder, J. A. 2002. Machine Learning: Intuition (Concept Learning) in Hybrid Neural Systems, NSA Technical Paper- CON-SP-0014-2002-06 Fort Meade, MD.
5. Crowder, J. A. 2010a. The Continuously Recombinant Genetic, Neural Fiber Network. Proceedings of the AIAA Infotech@Aerospace-2010, Atlanta, GA.
6. Crowder, J. A. 2010b. Anti-Terrorism Learning Advisory System (ATLAS): Operative Intelligent Information Agents for Intelligence Processing. Proceedings of the AIAA Infotech@Aerospace-2010, Atlanta, GA.
7. Crowder, J. A., 2010c. Flexible Object Architectures for Hybrid Neural Processing Systems. Proceedings of the 11th International Conference on Artificial Intelligence, Las Vegas, NV.
8. J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, and F. White. 2004. Revisiting the jdl data fusion model ii.
9. Carbone, J. N. 2010. A Framework for Enhancing Transdisciplinary Research Knowledge, Texas Tech University Press.
10. Crowder, J. A. and Crowder, J. A., and Carbone, J. N. 2011a. Recombinant Knowledge Relativity Threads for Contextual Knowledge Storage. Proceedings of the 12th International Conference on Artificial Intelligence, Las Vegas, NV.
11. Carbone, J. N. and Crowder, J. 2011b. Transdisciplinary Synthesis and Cognition Frameworks. Proceedings of the Society for Design and Process Science Conference 2011, Jeju Island, South Korea.
12. Crowder, J. and Friess S. 2012. Artificial Psychology: The Psychology of AI. Proceedings of the 3rd International Multi-Conference on Complexity, Informatics, and Cybernetics, Orlando, FL
13. Crowder, J. 2012a. Cognitive System Management: The Polymorphic, Evolutionary, Neural Learning and Processing Environment (PENLPE). Proceedings for the AIAA Infotech@Aerospace 2012 Conference, Garden Grove, CA.
14. Crowder, J. 2012b. The Artificial Cognitive Neural Framework. Proceedings for the AIAA Infotech@Aerospace 2012 Conference, Garden Grove, CA.
15. Crowder, J., Raskin, V., and Taylor, J. 2012. Autonomous Creation and Detection of Procedural Memory Scripts. Proceedings of the 13th Annual International Conference on Artificial Intelligence, Las Vegas, NV.
16. Raskin, V., Taylor, J. M., & Hempelmann, C. F. 2010. Ontological semantic technology for detecting insider threat and social engineering. New Security Paradigms Workshop, Concord, MA.
17. Rosenblatt F. 1962. Principles of Neurodynamics. Spartan Books.
18. Scally, L., Bonato M., and Crowder, J. 2011. Learning agents for Autonomous Space Asset Management. Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference, Maui, HI.
19. Taylor, J. M., & Raskin , V. 2011. Understanding the unknown: Unattested input processing in natural language, FUZZ-IEEE Conference, Taipei, Taiwan.
20. Dourish, P. 2004a. Where the action is: The foundations of embodied interaction. The MIT Press.
21. Dourish, P. 2004b. What we talk about when we talk about context. Personal and ubiquitous computing, vol. 8, pp. 19-30.
22. Torralba, A. 2003. Contextual priming for object detection. International Journal of Computer Vision, vol. 53, pp. 169-191.
23. Dey, A. 2001. Understanding and using context. Personal and ubiquitous computing, vol. 5, pp. 4-7.
24. Coutaz, J., Crowley, J., Dobson, S., and Garlan, D. 2005. Context is key. Communications of the ACM, vol. 48, pp. 53.
25. Winograd, T. 2001. Architectures for context. Human-Computer Interaction, vol. 16, pp. 401-419.
26. Hong, J. and Landay, J. 2001. An infrastructure approach to context-aware computing. Human-Computer Interaction, vol. 16, pp. 287-303.
27. Howard, N. and Qusaibaty, A. 2004. Network-centric information policy. Proceedings of the Second International Conference on Informatics and Systems.
28. Ejigu, D., Scuturici, M., and Brunie, L. 2008. Hybrid approach to collaborative context-aware service platform for pervasive computing. Journal of Computers, vol. 3, pp. 40.

Content Management in Digital Libraries

G. Concas, F. E. Pani, and S. Porru
Department of Electrics and Electronics Engineering, University of Cagliari, Cagliari, Italy

Abstract - *In recent years, the development of models to formalize knowledge has been studied and analysed. Many disciplines develop standardized formalization of knowledge, which domain experts can use to share information in the form of reusable knowledge. The purpose of this paper is to formalize knowledge through a mixed-iterative approach, applying a top-down and bottom-up analysis of the knowledge domain to represent. Our study is part of an industrial project being funded by the Autonomous Region of Sardinia, with the goal of implementing a web-based application with innovative functionalities in the field of semantic search and digital content management. We consider the case of Italian libraries, which have moved from holding mainly printed resources to a collection that includes also multimedia objects, such as music, databases, ebooks, audio sources, and websites, over the last decade.*

Keywords: Knowledge Management, top-down analysis, bottom-up analysis, taxonomy, multimedia standards.

1 Introduction

Italian libraries work independently, and at the same time are integrated into a cooperative system based on a national network, the SBN. SBN, the Italian libraries network, was founded in 1985, promoted by the Ministry of Cultural Heritage and Cultural Activities in cooperation with the Regions and the University, coordinated by the Central Institute for the Unified Catalogue of Italian Libraries and for Bibliographic Information (ICCU - Istituto Centrale per il Catalogo Unico).

Libraries in the SBN network are grouped into local Poles. Those Poles are in turn connected to the SBN Index system, the central node of the network, which contains the collective catalogue of publications. The shared catalogue makes their integration possible.

Over the last decade, SBN libraries have moved from a mainly paper-based heritage, to a hybrid heritage that combines printed resources and collections of multimedia objects of different types: music, audio, databases, e-books, audiobooks, Web sites. The evolution towards a digital library, a library that provides digital content loan services and browsing, is needed in order to meet the current and future needs of citizens, who are becoming more and more used to draw information from different types of digital content.

In this context, an issue that must be considered is the management of metadata associated with multimedia objects. The purpose of this paper is to formalize knowledge through a mixed-iterative approach, applying a top-down and a bottom-up analysis to the domain of interest.

Our study is part of a research project that is being funded by the Autonomous Region of Sardinia, and aimed at the implementation of a Web-based application intended for bibliographic cataloguing and library reference services. The application is going to have innovative features related to semantic search and digital content management; it involves also the creation of a social network. This Web application will make all cataloguing and statistics data produced by the libraries available as Open.

The paper is structured as follows: in Section Two we present an overview about the state of the art and in Section Three we propose our approach to multimedia objects management based on top-down and bottom-up analysis. In Section Four we present the case study and Section Five includes the conclusion and reasoning about the future evolution of the work.

2 Related work

Metadata standards are either application-specific or generic. Often, metadata standards provide information only for a particular type of multimedia object, or for a restricted set of multimedia objects, e.g., if we consider Exif standard [1], only for specific image and audio files. Moreover, it is difficult to integrate different metadata standards, because of the overlapping in functionality and their semantic ambiguity. Different standards are often not designed for a combined use. Many solutions have been proposed to provide a formal classification that could take into account the relationships between different multimedia metadata [2][3]. Ontologies based on the MPEG-7 [4][5] standard, like the one proposed by [6], the one proposed by [7], and the MPEG-7 Upper MDS [8] developed within the Harmony Project, all represented in OWL, are not suitable for an immediate use in the Italian digital library scenario, both for the higher emphasis placed on audio and video content than on other multimedia objects, and for the interoperability issues connected with the exploitation of the OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) [9]. The Multimedia Metadata Ontology (M3O) [10] is another solution to metadata standard integration issues. M3O is a modelling framework that targets the multimedia metadata standard integration issues by abstracting from existing standards. The alignment method used by the creators of M3O does not use machine learning approaches; instead, M3O makes use of a pure manual alignment, in order to ensure a high quality of standard integration. Creating an ontology is, usually, a complex task if compared to the process of creating a taxonomy [11][12][13][14]. A taxonomy generally is a much simpler formalization, that focuses on hierarchical relationships. Multimedia Metadata Ontology seems to lack in

simplicity, whereas a taxonomy-oriented, mixed-iterative approach can make it easy to achieve a more understandable formalization structure, both because of the emphasis put on the bottom-up phase, and of the use of a taxonomy instead of a complete ontology. Another relevant ontology to be considered is the Media Resource Ontology. Created by the W3C Media Annotation Working Group, it is an ontology based on the mapping definition of many different multimedia metadata standards, including Exif 2.2 [1], MPEG-7 [4][5], METS [15], NISO [16], and XMP [17]. It is mainly web-oriented, and, being structured following other standards, has the same drawbacks that multimedia ontologies previously mentioned have; that is, its construction does not rely on a bottom-up phase.
A remarkable effort to obtain a software-independent, and also hardware-independent, formalization, is the MAG standard [18]. The MAG schema is an application profile that interacts with other standards, namely Dublin Core [19], and NISO [16]. It is not like the modelling framework M3O, because the MAG schema defines a metadata taxonomy, so it is not as complex as an ontology, and it can achieve a higher degree of independence, both from application context, and from software and hardware. MAG metadata are specified through the XML format, in order to be compliant with the OAI-PMH standard. As an extensible standard, MAG could be a good starting point for the construction of a metadata taxonomy. Despite MAG being much easier to use than modelling frameworks like M3O, the approach used to build it lacks a bottom-up phase, on which the approach proposed in this paper is based. The bottom-up phase is fundamental to the construction of a classification that aims to be properly designed to be effectively usable, because it strives to take into account metadata that are actually used in the real world.
Another application profile, similar to MAG, is the PICO AP [20]. PICO AP, like MAG, is oriented to the exploitation of OAI-PMH, and its target is to ensure metadata harvesting functionalities also in the presence of different schemes, in addition to reaching a future-proof structure and supporting interoperability. Like MAG, PICO AP is an XML metadata schema which makes use of international standards. It is a DC application profile. PICO AP has been recently extended, via a specific encoding scheme, to support the encoding of metadata provided by ICCD (Istituto Centrale per il Catalogo e la Documentazione - Central Institute for Cataloguing and Documentation) cards.

3 Proposed approach

We want to represent knowledge through a mixed-iterative approach, applying a top-down and bottom-up analysis to the knowledge domain we need to investigate.

3.1 Top-down phase

When our knowledge or our expectations are influenced by perception, we refer to schema-driven or top-down (TD) elaboration. A schema is a model formerly created by our experience. General or abstract contents are placed at a higher level, while concrete details are placed at a lower level.
A TD elaboration happens whenever a higher level concept influences the interpretation of lower level information. Generally, the TD process is an information process based on former knowledge or acquired mental schemes; it allows us to make inferences: to “perceive” or “know” more than what can be found in data. The TD methodology starts, therefore, by identifying a target to reach, and then pinpoints the strategy to use in order to reach that target [21].
Our aim is to begin with a formalization of the reference knowledge (ontology, taxonomy, metadata schema) to start classifying the information on the reference domain.
The model could be, for instance, a formalization of one or more classifications of the same domain.

3.2 Bottom-up phase

In this phase, the knowledge to be represented is analysed by pinpointing, among the available information, what is needed, in order to define a reference terminology to describe the data.
We are going toanalyse the objects of interest in a domain, objects that contain the information of the domain itself; both information whose structure need to be extrapolated and the information in them are to be pinpointed.
One of the limits of this phase could lie in the creation of the KB, because each object can have a different structure and a different way of presenting the same information. Therefore, it will be necessary to pinpoint the present information of interest, defining and outlining it.

3.3 Iterations of phases

In this phase, we are going to try to reconcile the two representations of domain knowledge obtained in the previous phases.
Thus, we want to pinpoint, for each single metadata found in the TD phase, where the information can be found in the metadata representing the knowledge of each object (which, for us, represents the knowledge we want to represent, considering the semantic concept and not the way to represent it, absolutely subjective for every knowledge object).
Starting from this KB, further iterative refining can be made by re-analysing the information in different phases:
1) with a TD approach, checking if the information that is not represented by the chosen formalization can be formalized;
2) with a BU approach, analysing if some information of the Web sites can be connected to formalized items;
3) with the iterations of phases by which these concepts are reconciled.
This is obviously needed only for the information to be represented. The knowledge we want to represent is the one considered of interest by the users for the domain: for this reason, the most important pieces of information are chosen.
At the end of this analysis we are going to define a formalization, in form of ontologies, taxonomies, metadata schema, able to represent the knowledge of interest for this domain.
The final result of these phases will be a formalized knowledge able to be represented, reused and managed trough Knowledge Management Systems, where the knowledge of interest is available.

4 Case study

Through this study, we see how such a mixed-iterative approach made of top-down and bottom-up analyses applied to a knowledge domain could be efficient when formalizing knowledge.
Our real goal is to make knowledge manageable, shareable and reusable; we will focus our attention on information of interest in domain-specific knowledge.
We are proposing a work plan for a research project funded by the government of the Region of Sardinia, with the goal of implementing a Web-based application intended for the optimization of multimedia object metadata classification.
The basic starting concept is the definition of a KB: in our study, the knowledge-base is made by all kinds of multimedia objects that a digital library must manage: ebooks, audiobooks, music, websites, magazines, images.
Through a combined TD and BU approach, knowledge is extracted to define a common structure through a taxonomy, in order to classify and make the majority of such knowledge available.
This taxonomy allows for the definition of a reference knowledge to be managed in terms of really usable and interesting knowledge.
We are going to analyse the metadata standards used in multimedia contents management, and define a taxonomy to represent the semantics of these multimedia contents, so that in turn the metadata classification can give an unambiguous meaning.

4.1 Top-down analysis

We use standards such as, for example, Dublin Core, Exif and the XMP standard (a standard created by Adobe Systems Inc.) for processing and storing standardized and proprietary information relating to the contents of a file as a starting point of this domain.
We assume to use this approach because such standards allow to catalogue different aspects of multimedia content. With the TD analysis, it is possible to have a complete modelling of the domain of multimedia content properties, together with a uniform representation of the variety of associated metadata.
Below, we provide a brief introduction to the most relevant standards for the TD phase.

4.1.1 Dublin Core standard

Dublin Core is a standard for metadata that consists of a core of essential elements for the description of any digital material accessible via a computer network [22].
Becker et al. proposed a set of 15 basic elements also extended to sub-elements or qualifiers: each element is defined by using a set of 10 properties obtained by a standard ISO 11179 [23].
The main features of DC are as follows:
1) ease of use: the standard is aimed at specialized cataloguers which are not experts in cataloguing, as users;
2) semantic interoperability, which gives rise to a complex and precise data system the meaning of which has been agreed in advance, along with a value that allows the DC to be a standard for quality research in Internet;
3) flexibility, as it allows you to integrate and develop the data structure with different semantic meanings and a congenial application environment.

4.1.2 Exif standard

Exif (Exchangeable image file format) is a standard created by Japan Electronics and Information Technology Industries to specify the formats of digital systems handling image and sound files such as the ones used by digital cameras, scanners, and so on [1]. It is a standard supported by the main producers of digital cameras and it gives users the opportunity to supply photos with interchangeable information between imaging devices to improve processing and printing.
Exif offers a set of specific tags in itself, concerning shooting parameters and settings of the device at the time of capture. These cover a wide spectrum, including:
• time and date information, memorising the current date and time;
• camera settings, containing static information about the camera's model and producer, information about the orientation, aperture, shutter click speed, focal length, white balancing and ISO speed information for every image;
• information about shutter click's location, coming from a GPS receiver connected to the camera;
• information and descriptions about the copyrights.

4.1.3 XMP standard

The Adobe Extensible Metadata Platform (XMP) is a standard, created by Adobe Systems Inc., for processing and storing standardized and proprietary information relating to the contents of a file [17]. XMP standardizes the definition, creation, and processing of extensible metadata. Serialized XMP can be embedded into a significant number of popular file formats, without breaking their readability by non-XMP-aware applications. Embedding metadata avoids many problems that occur when metadata is stored separately. XMP is used in PDF, photography and photo editing applications. XMP encapsulates metadata inside the file, using RDF, a basic tool proposed by W3C to encode, exchange and reuse the structured metadata as proven by W3C.

Table 1: Formats supported by XMP

Image formats	Dynamic Media formats	Video Package formats	Adobe Applications formats	Markup formats	Docu-ment formats
DNG	ASF	AVCHD	INDD, INDT	HTML	PDF
GIF	AVI	P2	XML		PS, EPS
JPEG	FLV	HDV			UCF
JPEG-2000	MOV	XDCAM EX			
PNG	MP3	XDCAM, FAM			
TIFF	MPEG-2	AI			
	MPEG-4	PSD			
	SWD	INDD, INDT			
	WAV				

4.2 Bottom-up analysis

The BU analysis starts from the multimedia objects managed by digital libraries (ebooks, audiobooks, music, websites, magazines, images) and the information associated to them. These objects are varied and rich in many kinds of information. We intend to start to analyse those very different objects.

Primary, important information already emerges through object analysis and gathering: during a first skimming phase, the minimum, basic information necessary to appropriately describe our domain can be noticed. Then, important pieces of information are extrapolated by choosing fields or keywords which best represent the knowledge in order to create a knowledge base (KB).

The BU phase cannot be underestimated, as it is one of the main strengths in the proposed approach. This phase gives us the possibility to notice the truly relevant objects properties, those that are relevant for users, avoiding a biased evaluation. As we do not consider, at this time, the most widely used standards in the context of knowledge-base of interest, we can focus on the real objects, and not on the formalized representation given to us by standards. We can also focus on relevant object properties that widely used standards may not take into account.

After the analysis of information gathering, the following step is to create a classification that has to reflect, in the most faithful way, the structure of the knowledge in itself, respecting both its contents and hierarchy.

At the end of this analysis, we will use a taxonomy to define a formalization; the taxonomy needs to be able to represent the knowledge of interest.

5 Conclusions

The management of very complex knowledge is a big problem in Knowledge Management research; the main goal of our approach is in fact to reach a rational organisation of such large amounts of information.

Our approach is a simple process of applying a systematic analysis to capture, structure and manage knowledge. Our real goal is to make interesting knowledge available for sharing and reuse, and we focus our attention on interesting information in domain-specific knowledge.

We studied a process to identify existing formalizations and knowledge sources within the domain, paying attention to multimedia objects. Valuable knowledge is represented into explicit form through formalization and codification of information, in order to facilitate the availability of knowledge.

Our work will, therefore, allow to formalize the management of metadata associated with multimedia objects; the results of this study could be then used to derive an approach for the management of information related to multimedia objects, in order to implement functionalities for the search and classification of objects based on metadata associated with it.

6 References

[1] Technical Standardization Committee on AV IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif version 2.2. Published by Standard of Japan Electronics and Information Technology Industries Association, 2002. http://www.exif.org/Exif2-2.pdf

[2] María del Carmen Suárez-Figueroa, Mari Carmen, Ghislain Auguste Atemezing, Oscar Corcho. "The landscape of multimedia ontologies in the last decade". Multimedia tools and applications 62.2, pp. 377-399, 2013.

[3] Bernd Stadlhofer, Peter Salhofer, and Augustin Durlacher. "An Overview of Ontology Engineering Methodologies in the Context of Public Administration". SEMAPRO 2013, The Seventh International Conference on Advances in Semantic Processing. 2013, September 29 – October 03, 2013, Porto, Portugal, pp. 36-42 .

[4] P. Salembier, T. Sikora, B. S. Manjunath. “Introduction to MPEG-7: multimedia content description interface”. John Wiley & Sons, Inc., 2002.

[5] J. M. Martinez, R. Koenen, F. Pereira. “MPEG-7: the generic Multimedia Content Description Standard, part 1”. IEEE Multimedia, Vol. 9, pp. 78-87, 2002.

[6] R. García, O. Celma. “Semantic Integration and Retrieval of Multimedia Metadata”. Proceedings of the ISWC 2005 Workshop on Knowledge Markup and Semantic Annotation Semannot'2005, 185, page 69-80. CEUR Workshop Proceedings, 2005.

[7] C. Tsinaraki, P. Polydoros, S. Christodoulakis. “Interoperability support for Ontology-based Video Retrieval Applications”. Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR 2004), pp. 582–591, Dublin, Ireland, 21–23 July 2004.

[8] J. Hunter. “Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology”. International Semantic Web Working Symposium (SWWS), Stanford, July 30-August 1, 2001.

[9] C. Lagoze, H. Van de Sompel. “The making of the Open Archives Initiative protocol for metadata harvesting”. Library Hi Tech, 2003.

[10] A. Scherp, D. Eißing, C. Saathoff. “A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology”. International Journal of Semantic Computing, 6(01), 25-49, 2012.

[11] N. F. Noy, D. L. McGuinness. “Ontology Development 101: A Guide to Creating Your First Ontology”. Stanford Knowledge Systems, Laboratory Technical Report KSL-01-05, 2001.

[12] M. Hepp. “Ontologies: State of the Art, Business Potential, and Grand Challenges”. In: M. Hepp, P. De Leenheer, A. de Moor, Y. Sure, (Eds.), Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, ISBN 978-0-387-69899-1, Springer, pp. 3-22, 2007.

[13] T. Gruber. “Ontology”. In: L. Liu, M. T. Ãzsu, (Eds.), Encyclopedia of Database Systems, Springer- Verlag, 2008.

[14] Son Hoang Nguyen, Gobinda G. Chowdhury. "Designing and Engineering the Digital Library Ontology." Digital Libraries: Social Media and Community Networks. Springer International Publishing, 2013. 195-196.

[15] R. Gartner. “METS: Metadata Encoding and Transmission Standard”. JISC Techwatch report TSW, 02-05 , 2002.

[16] Denise M. Davis. “NISO Standard Z39.7 - The Evolution to a Data Dictionary for Library Metrics and Assessment Methods”. In Serials Review, Vol.30, Issue1, 2004. http://dx.doi.org/10.1016/j.serrev.2004.01.001.

[17] Adobe Systems Incorporated, Adobe XMP Specifications, additional properties, 2010. http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart2.pdf

[18] MAG, Comitato. Metadati amministrativi e gestionali: manuale utente, Elena Pierazzo (ed.), version 2.0. 1, marzo 2006. Roma: ICCU, 2006.

[19] Dublin Core Metadata Initiative, http://dublincore.org

[20] I. Buonazia, M. E. Masci, D. Merlitti. “The Project of the Italian Culture Portal and its Development. A Case Study: Designing a Dublin Core Application Profile for Interoperability and Open Distribution of Cultural Contents”. In: ELPUB, pp. 393-404, 2007.

[21] F. E. Pani, M. I. Lunesu, G. Concas, C. Stara, M. P. Tilocca. “Knowledge Formalization and Management in KMS”. In: Proceedings of the 4th International Conference on Knowledge Management and Information Sharing, KMIS 2012, Barcelona, Spain, 2012. ISBN: 978-989-8565-31-0.

[22] D. Hillmann. “Using Dublin Core”, 2005. Retrieved from: http://dublincore.org/documents/usageguide

[23] H. Becker, A. Chapman, A. Daviel, K. Kaye, M. Larsgaard, P. Miller, D. Nebert, A. Prout, M.P. Wolf. “Dublin Core element: Coverage”, 1997. Retrieved from: http://www.alexandria.ucsb.edu/public-documents/metadata/dc_coverage.html

Faces Recognition with Image Feature Weights and Least Mean Square Learning Approach

Wei-Li Fang, Ying-Kuei Yang and Jung-Kuei Pan
Dept. of Electrical Engineering, National Taiwan Uni. of Sci. & Technology, Taipei, Taiwan
Email: yingkyang@yahoo.com

Abstract - *Most of 2DPCA-enhanced approaches improve face recognition rate while at the expense of computation load. In this paper, an approach is proposed to greatly improve face recognition rate with slightly increased computation load. In this approach, the 2DPCA is applied against a face image to extract important image features for selection. A weight is then assigned to each of selected image features according to the feature's importance to face recognition. The least mean square (LMS) algorithm is further applied to optimize the feature weights based on the recognition error rate during learning process in order to improve face recognition performance. The experiments have been conducted against ORL face image database to make performance comparisons among several better-known approaches, and the experimental results have demonstrated that the proposed approach not only has excellent face recognition rate of 99% but also requires only slightly higher computation load than 2DPCA, making the approach more practical to real face recognition applications..*

Keywords: face recognition, feature extraction, principle component analysis, least mean square, weight assignment, steepest decent algorithm.

1 Introduction

Face recognition in image processing has been significantly important because it can be applied in human life efficaciously. Research areas include building/store access control, suspect identification, security and surveillance [1]-[11].

Seceral algorithms have been proposed in face recognition. The best ones should be those that try not only to reduce computation cost but also to increase recognition rate [13][14]. Based on this viewpoint, principal component analysis (PCA) [15] has become a popular feature extraction algorithm in recent decades.

After PCA was proposed, Yang *et al.* [13] proposed the so-called two-dimensional principal component (2DPCA) algorithm aiming for better feature extraction of face images. The 2DPCA has achieved the goal of increasing recognition rate and reducing computation cost simultaneously [13]. Because 2DPCA has such good performance, various face recognition algorithms based on 2DPCA had been proposed and enhanced. For instance, the approach of "Two-directional two-dimensional PCA ($(2D)^2PCA$)" proposed by Zhang *et al.* [17] is to process a face image from transverse and longitudinal axis respectively and then perform the recognition by analyzing their shortest dimension. Low computation cost is the advantage of this approach. Unfortunately, its improvement on recognition rate is not ubiquitous in relatively large scale of training samples [16]. Sanguansat *et al.* [18] proposed the approach of "Two-dimensional principal component combined two-dimensional Linear discriminant analysis (2DPCA&2DLDA)" [18] to face recognition applications. Although this approach solves the small sample size problem, its computation cost is high due to the composition of 2DPCA and 2DLDA. Meng *et al.* [19] proposed the combination of 2DPCA with self-defined volume measure to perform feature extraction by 2DPCA first and then conduct classification by computing the distances of matrix volumes. This approach is more suitable to process applications with high dimensional data. Wang *et al.* [20] proposed "probabilistic two-dimensional principal component analysis" that combines 2DPCA with Gaussian distribution concept to mitigate the noise influence in face image recognition. Kim *et al.* [21] proposed "fusion method based on bidirectional 2DPCA" that reduces dimensions of both row and column vectors before performing face recognition procedure. It does increase recognition rate, but at the expense of high computation cost [22].

Aforesaid face recognition algorithms all have pros and cons. The algorithm "2DPCA" is especially designed for face image data, so the recognition performance is better than using traditional PCA. In this paper, an approach is proposed hoping to achieve the goal of increasing the recognition rate while not at expense of computation cost in face image recognition. This approach incorporates weights in projected feature vectors of 2DPCA and uses least mean square (LMS) algorithm to optimize the weights based on the recognition error rate during learning process in order to achieve better face recognition performance.

2 The least mean square-two dimensional principal component analysis

2.1 Two-dimensional principal component analysis (2DPCA)

The 2DPCA approach by Yang *et al.* [13] in 2004 is proposed particularly for two dimensional image data. Suppose there is an image data set $Z=\{\mathbf{A}_1, \mathbf{A}_2,\ldots, \mathbf{A}_N\}$ with N images, and the dimension of every image is $n \times n$. The covariance matrix of the image data set is computed by Eq. (1) and the average value of the data set is computed by Eq. (2).

$$\mathbf{R} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{A}_i - \overline{\mathbf{A}})(\mathbf{A}_i - \overline{\mathbf{A}})^T \quad (1)$$

$$\overline{\mathbf{A}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{A}_i \quad (2)$$

where $\mathbf{A}_i$ is an image in the data set, $\mathbf{R}$ is covariance matrix, and $\overline{\mathbf{A}}$ is data average.

After eigen-decomposition is performed for covariance matrix, k eigenvectors corresponding to the k biggest eigenvalues are selected. These eigenvectors are the projection vectors of the original image data set and the features of the image can therefore be extracted from those projection vectors as shown in Eq. (3).

$$\mathbf{Y}_i = \mathbf{A}\,\mathbf{X}_i \qquad i=1,2,\ldots,k \quad (3)$$

where $\mathbf{Y}_i$ are projected feature vectors, $\mathbf{X}_i$ means eigenvectors. Suppose there are k biggest eigenvalues being selected, then a feature vector set $\mathbf{B}=[\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k]$ in descending order of eigenvalues can be obtained and these projected feature vectors are the resultant principal components of an original image data $\mathbf{A}$ by 2DPCA.

Because 2DPCA processes a 2-dimensional face image directly, it can get better result of feature extraction. On the contrary, the conventional PCA needs to transform an image into one-dimensional data and therefore loses some feature information. Consequently, the recognition rate by 2DPCA is better than conventional PCA for 2-dimensional face images.

2.2 Least mean square algorithm (LMS)

Least mean square (LMS) is an adaptive filter algorithm in signal process [23], and it is applied in many engineering fields. Its input signals $\boldsymbol{u}(n)$ are computed by transversal adaptive filter to result in the output $y(n)$. The desired signal is $d(n)$ and $e(n)$ is the difference between actual output y(n) and desired output $d(n)$. After training by iteration process, $e(n)$ becomes smaller and smaller meaning the adaptive filter is closer to the ideal state.

The main essence of LMS is to make the error rate $e(n)$ as smaller value as possible. Hence, the cost function is defined as the expected value of squaring error rate, as shown in Eq. (4).

$$J(n) = E[e^2(n)] \quad (4)$$

In Eq. (4), the square operation is needed to avoid the problem caused by different sign characteristics because the error rate could be a either positive or negative value. The steepest decent algorithm[23] is then performed against the cost function to make the resultant error rate as small as possible. This operation process is shown as Eq. (5).

$$\hat{\mathbf{w}}(n+1) = \hat{\mathbf{w}}(n) + \mu\mathbf{u}(n)e(n) \quad (5)$$

Eq. (5) represents the process of adjusting weights by iteration operation. The symbol μ is step size. The learning process is repeated until the error rate has reached a pre-set satisfactory value.

2.3 The integration of least mean square with two-dimensional principal component analysis

The feature extraction algorithm 2DPCA has good performance in face recognition. Important features that are represented by projected feature vectors are selected during the process of eigen-decomposition. One projected feature vector represents one extracted feature. The projected feature vector that corresponds to the biggest eigenvalue represents the most important feature; the one that corresponds to second biggest eigenvalue represents the second important feature; and so on. After eigen-decomposition, the projected feature vectors are arranged in a row in the descending order of feature importance.

For 2DPCA and most of its extensions, every projected feature vector has equal weight. This is not a good idea in terms of improving recognition performance since the importance of each projected feature vector, meaning each feature, is different. Rather, the weight assigned to a projected feature vector should be related to the importance of the feature to that a projected feature vector corresponds. That is, the projected feature vector corresponding to the biggest eigenvalue should have highest weight during the process of face recognition.

Although methods have been proposed to assign different weights to projected feature vectors, most of them decide these weights based on trial-and-error process which is not only time-consuming but inefficient. In this paper, an approach is proposed by integrating least mean square with two-dimensional principal component analysis in order to efficiently obtain proper weight for each of selected features hoping to improve face recognition performance. The proposed approach uses LMS to dynamically adjust the weights of projected feature vectors associated with image features. Weights are adjusted based on the feedback of error rate calculated by each iteration.

Fig. 1 shows the proposed system structure based on the concept of an adaptive filter. In Fig. 1, $\boldsymbol{u}(n)$ is the projected feature vectors generated by 2DPCA, and is multiplied by weight matrix to get the output $y(n)$ through the transversal adaptive filter. The error rate e(n) is then calculated by nearest neighbor rule (NNR) and then further used by LMS inside the weight-control mechanism to dynamically adjust the weights of projected feature vectors. The weight matrix is initially set

as $\hat{\mathbf{w}}(n)_{N\times m\times d}$ that has dimension N×m×d and value 1 in all matrix elements, where *n* means *n*-th iteration starting from initial value 1. The N means data amount and m×d is the dimension of every data.

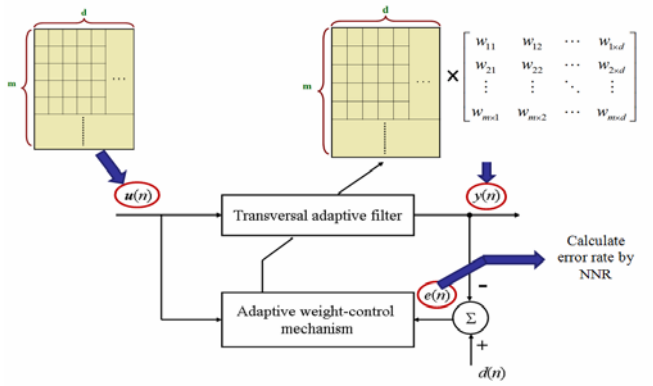


Fig 1: The system structure

To simplify the computation, the error rate of face recognition system is calculated by the nearest neighbor rule that is based on Euclidean distance shown in Eq. (6).

$$d = \left\| \mathbf{V} - \mathbf{P} \right\|_2 \tag{6}$$

The symbols **V** and **P** are vectors, and d is Euclidean distance. The operation of Eq. (6) computes the norm of **V**-**P**. Suppose **V**=(v_1, v_2, v_3) and **P**=(p_1, p_2, p_3), the norm of the two vectors is obtained as Eq. (7).

$$\left\| \mathbf{V} - \mathbf{P} \right\|_2 = \sqrt{(v_1 - p_1)^2 + (v_2 - p_2)^2 + (v_3 - p_3)^2} \tag{7}$$

After calculating the error rate by NNR, it is used by LMS iteration to adjust the feature weights as shown in Eq. (5). A threshold value is set for the error rate to avoid infinite iteration.

The face recognition rate is calculated as below. Suppose there are N face images, represented as $\mathbf{B}_1$, $\mathbf{B}_2$,…, $\mathbf{B}_N$, and each image is represented by a projected feature vector, such as [$\mathbf{Y}_1^1$, $\mathbf{Y}_2^1$,…, $\mathbf{Y}_d^1$] for $\mathbf{B}_1$ with m×d dimension. The classes of these N images are already known. Suppose a class-unknown image $\mathbf{T}_k$=[$\mathbf{T}_1$, $\mathbf{T}_2$,…, $\mathbf{T}_d$] is to be recognized against these N face images. The computation process is shown in Eq. (7).

$$d(\mathbf{B}, \mathbf{T}) = \sum_{k=1}^{d} \left\| \mathbf{B}_k - \mathbf{T}_k \right\|_2 \tag{8}$$

where d is the calculated distance by NNR between the two images. The class of $\mathbf{T}_k$ is classified as the class of $\mathbf{B}_k$ if these two have minimum distance *d* in Eq. (8). The face recognition rate can therefore be obtained after classifying all N face images.

3 Experiments and analysis

The ORL database [24] is a well-known face image database and is used in this paper for experiments. There are 40 individual faces in ORL database. Each individual face has 10 different images making totally 400 face images in the database. The images were taken with a tolerance of some tilting and rotation of the face for up to 20 degrees [13][24]. In ORL database, all images are grayscale with dimension of 112×92. The pixel value range is 0~255.

Among the 10 different images of each individual face, 5 face images are selected as training data and the rest of 5 face images are used as testing data, making totally 200 images for training data and 200 images for testing data.

Fig 2 shows the error rate values during the first 300 LMS iterations respectively in the conducted experiment. In Fig. 2, the lowest error rate takes place at around 52th LMS iteration. It can also be observed in Fig 2 that the error rate is stabilized at certain value after around 135th LMS iteration, which means the feature weights have been learned to be the most appropriate values.

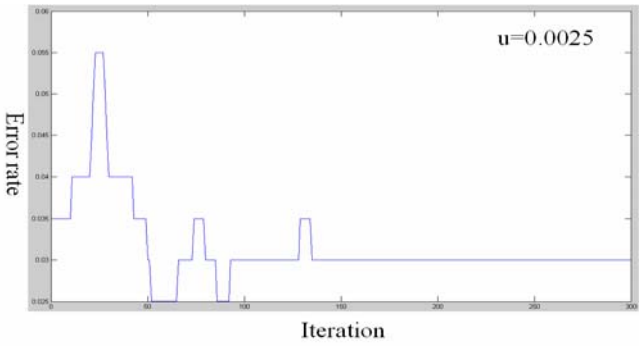


Fig. 2: Error rate during LMS iterations (300 times)

To see the improvement on face recognition during the LMS learning procedure, the face recognition rate is performed after each LMS iteration. The experimental result is shown in Fig. 3. The face recognition rate reaches the best value of 99% starting from around 52th LMS iteration in Fig. 3, which coincides with Fig. 2 that shows the lowest error rate takes places at 52th LMS iteration. Since then, the face recognition rate maintains at value 99% as the feature weights have been adjusted to appropriate values by the LMS learning procedure at this moment.

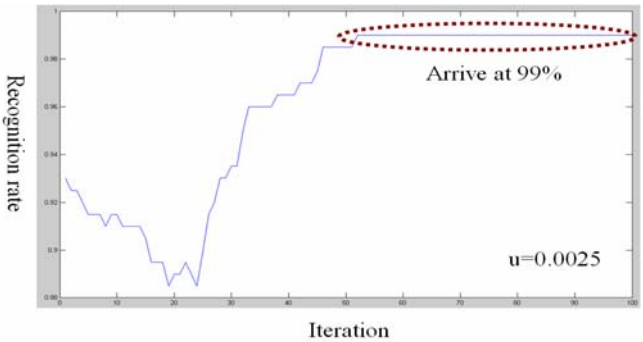


Fig. 3: Face recognition rate during LMS iterations

In Table 1, the experimental result of the proposed LMS-2DPCA in this paper is compared against some other methods which are enhancements from 2DPCA. The experiments are conducted against ORL database for all the methods indicated in Table 1. The table shows that the proposed approach has the best face recognition rate of 99% while the computation load is only normal. Although method 1 [17] has slightly lower computation load than the proposed approach, its face recognition rate is much lower. The good recognition performance of the proposed approach comes from the good adjustment to the image feature weights by LMS learning procedure. The normal computation load is contributed by the LMS' simple algorithm, making the whole computation cost is only slightly more than pure 2DPCA

Table 1: Performance comparison between the proposed approach and other methods

Method number	Method	Recognition rate	Computation cost
1	$(2D)^2$PCA [17]	90.5%	normal
2	2DPCA+Fusion method based on bidirectional [21]	92.5%	normal
3	2DPCA+2DLDA [18]	93.5%	normal
4	2DPCA+Kernel [25]	94.58%	high
5	OP-SRC[26]	95.00%	high
6	RC2DPCA[27]	96.65%	Very high
7	2DPCA+Feature fusion approach [28]	98.1%	very high
8	MMDA[29]	98.31%	Very high
9	**Proposed approach**	99%	normal

4 Conclusions

The 2DPCA is a good approach for 2-dimensional face image recognition. Although enhanced approaches based on 2DPCA have been proposed, most are either too time-consuming or no much improvement to face recognition. The 2DPCA treats all selected image features same weight in terms of recognition. However, the importance or influence to face recognition from each image feature is different from one another, meaning each image feature should be assigned an appropriate weight according to its influence to face recognition. Therefore, this paper proposes an approach that integrates 2DPCA with LMS learning procedure. The 2DPCA is applied against a face image to extract important image features for selection. Then the LMS learning procedure is applied to the training samples to assign the most appropriate weight to each of selected image features hoping to increase the face recognition rate. Because the goal is to make the face recognition error rate as small as possible, the image feature weights are adjusted based on the feedback of face recognition error amount by LMS iterations. Due to the simple algorithm, the additional computation cost required to run LMS learning procedure is only to a small extent of slightly more than pure 2DPCA. The experiments conducted in this paper has shown that the proposed approach not only has excellent face recognition rate of 99% but also requires only slightly higher computation load than 2DPCA, making the approach more practical to real face recognition applications.

5 References

[1] Q. Liu, X. Tang, H. Lu and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 1081–1085, Jul. 2006.

[2] W. Zheng, X. Zhou, C. Zou and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.

[3] X. Tan, S. Chen, Z. H. Zhou and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 875–886, Jul. 2005.

[4] P. Melin, O. Mendoza and O. Castillo, "Face recognition with an improved interval type-2 fuzzy logic Sugeno integral and modular neural networks," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 5, pp. 1001–1012, Sep. 2011.

[5] N. Sudha, A. R. Mohan and P. K. Meher, "A self-configurable systolic architecture for face recognition system based on principal component neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 8, pp. 1071–1084, Aug. 2011.

[6] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[7] N. S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.

[8] J. Y. Choi, Y. M. Ro and K. N. Plataniotis, "Color local texture features for
color face recognition," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1366–1380, Mar. 2012.

[9] H. Chen, Y. Y. Tang, B. Fang and J. Wen, "Illumination invariant face recognition using FABEMD decomposition with detail measure weight," *IJPRAI*, vol. 25, pp. 1261- 1273, 2011.

[10] H. Yu, J. J. Zhang and X. Yang, "Tensor-based feature representation with application to multimodal face recognition," *IJPRAI* , vol. 25, pp. 1197-1217, 2011.

[11] G. Chiachia, A. N. Marana, T. Ruf and A. C. Ernst,

"Histograms: A simple feature extraction and matching approach for face recognition," *IJPRAI*, vol. 25, pp. 1337-1348, 2011.

[12] Rabia Jafri and Hamid R. Arabnia, "A Survey of Face Recognition Techniques", *Journal of Information Processing Systems*, pp. 41-68, Vol.5, No.2, June 2009

[13] J. Yang, D. Zhang, A. F. Frangi and J. Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*., vol. 26, no. 1, pp. 131-137, Jan. 2004.

[14] J. Lu, X. Yuan and T. Yahagi, "A method of face recognition based on fuzzy c-means clustering and associated sub-NNs," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, Jan. 2007

[15] L. Sirovich and M. Kirby, "Low-dimensional procedure for characterization of human faces," *J. Optical Soc. Am.*, vol. 4, pp. 519-524, 1987.

[16] W. H. Yang and D. Q. Dai, "Two-dimensional maximum margin feature extraction for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 1002-1012, Aug. 2009.

[17] D. Zhang and Z. H. Zhoua, "$(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, pp. 224–231, Jun. 2005.

[18] P. Sanguansat, W. Asdornwised, S. Jitapunkul and S. Marukatat, "Tow-dimensional linear discriminant analysis of principle component vectors for face recognition," *ICASSP 2006*, pp. 345-348, May. 2006.

[19] J. Meng and W. Zhang, "Volume measure in 2DPCA-based face recognition," *Pattern Recognition Lett.*, vol. 28, pp. 1203–1208, Jan. 2007.

[20] H. Wang, S. Chen, Z. Hu and B. Luo, "Probabilistic two-dimensional principal component analysis and its mixture model for face recognition," *Springer Neural Comput & Applic*, vol. 17, pp. 541–547, 2008.

[21] Y. G. Kim, Y. J. Song, U. D. Chang, D. W. Kim, T. S. Yun and J. H. Ahn, "Face recognition using a fusion method based on bidirectional 2DPCA," *Applied Mathematics and Computation.*, vol. 205, pp. 601–607, 2008.

[22] Y. Qi and J. Zhang, "$(2D)^2$PCALDA: An efficient approach for face recognition," *Applied Mathematics and Computation.*, vol. 213, no. 1, pp. 1-7, Jul. 2009.

[23] S. Haykin, *Adaptive Filter Theory*, 4rd Edition, Prentice-Hall, 2001.

[24] "The ORL face database", http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.htm

[25] N. Sun, H. X. Wang, Z. H. Ji, C. R. Zou and L. Zhao, "An efficient algorithm for kernel two-dimensional principal component analysis," *Neural Comput & Applic.*, 17, pp. 59-64, 2008.

[26] C. Y. Lu and D. S. Huang, "Optimized projections for sparse representation based classification," *Neurocomputing*, vol. 113, pp. 213-219, Mar, 2013

[27] W. Yang, C. Sun and K. Ricanek, "Sequential row–column 2DPCA for face recognition," *Neural Comput & Applic.*, vol. 21, pp. 1729-1735, 2012.

[28] Y. Xu, D. Zhang, J. Yang and J. Y. Yang, "An approach for directly extracting features from matrix data and its application in face recognition," *Neurocomputing*, 71, pp. 1857-1865, Feb, 2008.

[29] W. Yang, C. Sun and L. Zhang, "A multi-manifold discriminant analysis method for image feature extraction," *Pattern Recognition*, vol. 44, pp. 1649-1657, Feb. 2011.

A Knowledge Based Selection Framework for Cloud Services

Gülfem Isiklar Alptekin[1] **and S. Emre Alptekin**[2]
[1]Computer Engineering, Galatasaray University, İstanbul, Turkey
[2]Industrial Engineering, Galatasaray University, İstanbul, Turkey

Abstract - *Cloud computing is a scalable services consumption and delivery platform where resources (computational processing power, storage, etc.) are retrieved from the network from anywhere in the world. The inherent complexity and elasticity of the cloud platform products makes their selection a difficult decision for their prospective customers. This paper proposes a multi-criteria based decision support tool which incorporates customer expectations and product attributes and their interrelationships into the decision process. Based on this knowledge the customers are able to rank various alternatives. The proposed knowledge based decision framework is based on quality function deployment and analytic network process. The applicability of the proposed methodology is demonstrated via a real life scenario.*

Keywords: Cloud computing; service selection; quality function deployment.

1 Introduction

Although there are many definitions of cloud computing, the most cited is the one of National Institute of Standards and Technology (NIST) [1]: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability." NIST, Hoefer et al. [2] and Buyya et al. [3] differentiate cloud services and creates three classes: SaaS (Software as a Service), PaaS (Platform as a Service) and IaaS (Infrastructure as a Service). In this work, we will concentrate on evaluating services provided as IaaS.

In cloud market, there are various service providers offering diverse range of configurations to satisfy different customer requirements. The decision of the appropriate configuration is usually a challenging task for an average customer. In this paper, we aim at offering a decision support framework for customers in choosing the most suitable product. Doing so, we have used Quality Function Deployment (QFD) approach, a common tool which simply intends to analyze customers' needs (CNs) to guarantee satisfaction. The application of QFD starts with the development of the house of quality (HOQ). HOQ uses customer feedbacks expressed as needs for input and tries to transform this knowledge into product attributes, which represent the characteristics of a product from a technical view. During the transformation process of CNs into the product technical requirements (PTRs), relationships between CNs and PTRs, and correlation between PTRs need to be determined. This transformation enables to obtain the weights for PTRs, which represent the most important characteristics to concentrate on, in order to satisfy customers. Herein, for obtaining the weights, we have utilized Analytic Network Process (ANP) [4]. ANP is a generalization of Saaty's Analytic Hierarchy Process (AHP), which is one of the most widely used multi-criteria decision support tools [5]. Most of real life decision problems cannot be structured as a hierarchy, since they include interaction and dependence of higher level elements in a hierarchy on lower level elements. Therefore, the hierarchy becomes more like a network. On this context, ANP and its supermatrix technique can be considered as an extension of AHP that can handle a more complex decision structure [6] [7] as the ANP framework has the flexibility to consider more complex interrelationships (outerdependence) among different elements.

In this work, we have used the final weights for PTRs and combined them with a competitive analysis. This analysis includes different service providers' performance in terms of technical attributes. A simple weighted average calculation combining performance values with weights enabled us to rank service providers. The applicability of the proposed approach is demonstrated via the case provided and evaluated by Garg et al. [8].

The paper is structured as follows. Section 2 describes related literature. The methodologies used in the approach are given in Section 3, while Section 4 presents step by step explanation of the research framework. Section 5 reveals the results and the concluding remarks of the case study and future works are given in Section 6.

2 Related Work

Although there are lots of applications of QFD with ANP approach, we have concentrated on the applications in cloud computing field. In one of the recent studies, the authors have proposed a model of cloud service selection by aggregating the information from both the feedback from

cloud users and objective performance analysis from a trusted third party [9]. They have used a fuzzy simple additive weighting system in order to select the best cloud service. Another work made use of the AHP approach to select most appropriate SaaS product in terms of five criteria (i.e., functionality, architecture, usability, vendor reputation and cost) [10]. Their research is mainly based on subjective assessment. The integrated AHP and fuzzy technique for order preference by similarity to an ideal solution (TOPSIS) approach was used in another cloud service comparison approach [11]. They proposed a standardization process of the performance attributes, but it is not sufficient to deal with real life's more complex cloud services. In another article, fuzzy TOPSIS approach is utilized to help service consumers and providers to analyze available web services with fuzzy opinions [12]. The authors ranked available alternative web services according to group preference. In their work, Ranjan et al. [13] presented a framework (called CloudGenius) which automates the decision-making process based on a model and factors specifically for Web server migration to the Cloud. They used AHP to automate the selection process based on a model, factors, and QoS parameters related to an application. Assuming that each individual parameter affects the service selection process, and its impact on overall ranking depends on its priority in the overall selection process, Garg et al. [11] proposed an AHP based ranking mechanism to solve the problem of assigning weights to features considering interdependence between them, thus providing a much-needed quantitative basis for ranking of Cloud services. In their paper, Ergu et al. [14] proposed a framework for SaaS software packages evaluation and selection by combining the virtual team (VT) and the BOCR (benefits, opportunities, costs, and risks) of the analytic network process (ANP). They attempted to solve the complex ANP model by decomposing the tasks to different parts, and performed by benefits virtual team (BVT), opportunities virtual team (O-VT), costs virtual team (C-VT), and risks virtual team (R-VT) separately.

The main contribution of this study over previous cloud service selection methodologies is that the proposed methodology enables to incorporate customer feedback in more complete and systematic way. The interrelationships between customer feedback and also the interrelationships between the technical attributes could be analyzed and used in the selection process. Hence, customer attributes with little or no meaning to customer can be identified and more importance to aspects meaning a lot can be given. The decision framework presented in this paper has advantages in comparison to the previously proposed analytical approaches such as the commonly applied analytic hierarchy process (AHP). AHP assumes that preferential independence of the technical attributes hold; however, this assumption generally does not hold in real-world applications.

3 Methodologies

3.1 Quality Function Deployment (QFD)

In literature QFD methodology is usually seen as a strategic knowledge management tool that tries to incorporate customer feedback into the product/service development process, which gathers knowledge from different functions of the organization and aims a successful product/service in terms of profitability and customer satisfaction. QFD provides a framework that deals with the knowledge gathered from different sources and combine them in a systematic and meaningful way.

QFD methodology is usually applied using several steps generally referred as matrices that are deployed as means for information transformation requiring different inputs from different functions and connecting them in a way so that each step's output simply becomes the input for the following one. There are many studies in literature that brought different aspects of QFD together and presented them as a literature survey. Chan and Wu's study is among the most addressed studies [15]. QFD methodology is used for knowledge transformation and requires different so called matrices. The first of these matrices is usually named as house of quality (HOQ). A demonstrative house of quality matrix is shown in Figure 1. As depicted in the figure, house of quality comprises eight elements:

(1) Customer needs (CNs) (WHATs). As the initial input for QFD, they are the essential knowledge block that should be incorporated to the development process. They are expressed deliberately in customers' own phrases, so that the main knowledge is retained in its original form.

(2) Product technical requirements (PTRs) (HOWs). They embody the knowledge of product/service in terms of technical attributes. They are used to achieve the goals set by CNs by providing alternative means to systematically change the product characteristics.

(3) Relative importance of the CNs. The knowledge CNs provide, is usually too diversified to deal with simultaneously. Hence, at this stage most important CNs have to be identified in order to increase the probability of a greater customer satisfaction. In most cases, organizations have to deal with conflicting demands and this circumstance usually means that an important trade-off has to be made.

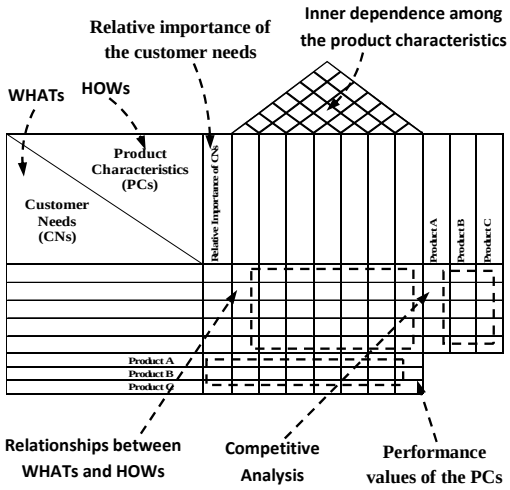


Fig. 1. House of quality

(4) Relationships between WHATs and HOWs. This relationship element is usually placed in the body of the house of quality and denotes the knowledge to what extent each PTR affects each CN. This step is very important as the transformation of different information occurs. The expected result of this stage is the importance of CNs presented in terms of PTRs.

(5) Inner dependencies among the CNs. The diversification of CNs is a difficult matter that should be solved diligently. At this stage we simply try to determine the interaction among the CNs. The resulting information could be used to measure how much and whether or not CNs support each other.

(6) Inner dependencies among the PCs. The inner dependencies among PCs are placed in the roof of house of quality and similar to the inner dependencies between CNs, they are used to measure to what extent a change in one feature may affect another.

(7) Competitive analysis. At this stage the benchmarking process tries to indicate improvement directions necessary to achieve total customer satisfaction by including competitors' performance into the decision process.

(8) Overall priorities and performance values of PCs. The performance values of PCs and the PCs' final ranking is usually used to establish a final ranking of PTRs.

3.2 Analytic Network Process (ANP)

ANP has its origins in the widely used multi-criteria decision making tool, the Analytic Hierarchy Process (AHP). AHP simply decomposes a problem into several levels in such a way that they form a hierarchy, where each element is supposed to be independent [16]. AHP incorporates both qualitative and quantitative approaches to a decision problem [17]. But AHP cannot deal with interconnections and innerdependencies among decision factors at the same level [18]. In order to deal with this shortcoming, ANP is developed by replacing hierarchies with networks and is used as an effective tool in those cases where the interactions among the elements of a system form a network structure [18].

The interactions among the elements in ANP are evaluated using pairwise comparisons. Accordingly, a supermatrix is obtained by these priority vectors, which is a matrix of influence among the elements. It is raised to limiting powers to calculate the overall priorities, and thus the cumulative influence of each element on every other element is obtained [19]. The supermatrix of a hierarchy with three levels is as follows:

$$\mathbf{W} = \begin{array}{c} \\ \text{Goal(G)} \\ \text{Criteria(C)} \\ \text{Alternatives(A)} \end{array} \begin{array}{c} \begin{array}{ccc} \text{G} & \text{C} & \text{A} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ \mathbf{W}_{21} & 0 & 0 \\ 0 & \mathbf{W}_{32} & \mathbf{I} \end{pmatrix} \end{array}$$

where $\mathbf{W}_{21}$ is a vector that represents the impact of the goal on the criteria, $\mathbf{W}_{32}$ is a matrix that represents the impact of the criteria on each of the alternatives, and I is the identity matrix. When a network consists of only two clusters apart from the goal, namely criteria and alternatives, the matrix manipulation approach proposed by [20] can be employed to deal with dependence of the elements of a system. This approach, which will be described in section 4, is based on [21]. It is used herein to incorporate the dependencies inherent in QFD process into the analysis. 1. The end users announce their service requests to all the application service providers in their region,

4 Research Framework

The proposed decision framework evaluates customer satisfaction by rating different service provider products based on customer feedback and technical attribute performance values. The proposed methodology is based on the work of Karsak et al. [21]. The main difference compared to Karsak et al. is that their methodology is suggesting improvement directions for technical attributes whereas our methodology is able to rate different products based on customer feedback and competitive analysis results. In this work, we have combined QFD methodology with ANP. The supermatrix representation required by ANP is constructed as follows:

$$\mathbf{W} = \begin{array}{c} \\ \text{Goal (G)} \\ \text{Criteria (C)} \\ \text{Alternatives (A)} \end{array} \begin{array}{c} \begin{array}{ccc} \text{G} & \text{C} & \text{A} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ \mathbf{w}_1 & \mathbf{W}_3 & 0 \\ 0 & \mathbf{W}_2 & \mathbf{W}_4 \end{pmatrix} \end{array}$$

where $\mathbf{w_1}$ is a vector representing the impact of the goal, namely a product/service that will satisfy the customers, $\mathbf{W_2}$ is a matrix that denotes the impact of the CNs on each of the PTRs, $\mathbf{W_3}$ and $\mathbf{W_4}$ are the matrices that represent the inner

dependencies of the CNs and PTRs, respectively.

When a network consists of only two clusters apart from the goal, namely criteria and alternatives, the matrix manipulation approach proposed by [20] can be employed to deal with dependence of the system elements. Thus, the interdependent priorities of the CNs ($\mathbf{w_C}$) are computed by multiplying $\mathbf{W_3}$ by $\mathbf{w_1}$, and similarly the interdependent priorities of the PTRs ($\mathbf{W_A}$) are obtained by multiplying $\mathbf{W_4}$ by $\mathbf{W_2}$. Overall priorities of the PTRs ($\mathbf{W_{ANP}}$) are obtained by multiplying $\mathbf{W_A}$ and $\mathbf{W_C}$. Next, the performance values for each cloud service providers' product is obtained in terms of PTR values. These performance levels are then normalized in order to overcome the problem of commensurability. The obtained normalized performance values are combined with the weights of each PTR. The result is used to rank the products. The main steps and knowledge processed in each are summarized as follows:

Step 1. QFD process determines the CNs, which are customers' perceptions and linguistic assessments in respect to the product/service. The PTRs, the tools of the company used to satisfy these CNs, are also identified in this step. The CNs and the PTRs used in this study are based on the work of Garg et al. [8]. Our main motivation for this choice was that they established a comprehensive list of attributes which could be categorized as CNs and PTRs. We used their attributes and classified quality related attributes as CNs and performance related attributes as PTRs. But, the list of CNs and also PTRs could be extended based on Cloud Service Measurement Index Consortium's measurement indexes [22] and also ISO/IEC 25010:2011 standard which defines a product quality model for software.

Step 2. As mentioned in previous section, the most important CNs have to be determined in order to make the necessary tradeoffs. Herein, we have used pairwise comparisons as suggested by ANP. As a result, we have obtained $\mathbf{w_1}$.

Step 3. In this step, assuming that there is not any dependence among PTRs, the degrees of relative importance of PTRs with respect to each CN are identified. As a result, we have obtained $\mathbf{W_2}$.

Step 4. It is not possible to assume that CNs are independent in real life scenarios. Therefore, we have used ANP to determine the interdependence among CNs. Once again, we have used pairwise comparisons and have obtained $\mathbf{W_3}$.

Step 5. Similarly, as PTRs may affect each other, we have determined interdependencies among them. The resulting matrix is $\mathbf{W_4}$.

Step 6. At this stage we transformed customer requirements into measurable technical requirements. For this transformation, we have initially calculated interdependent priorities of CNs ($\mathbf{W_C}$) and also interdependent priorities of PTRs ($\mathbf{W_A}$) and have combined them to obtain the overall priorities of PTRs.

Step 7. The performance values of each cloud service provider for each PTR are evaluated in this step. The obtained performance values are normalized and using simple weighted average formulation final ratings for cloud products are calculated. The ratings are used for ranking the products.

5 Case Study

As a demonstrative example, we have used the data provided in the work of Garg et al. [8]. They also aimed to select the best cloud service provider using real world data. They rated Amazon EC2, Windows Azure and Rackspace in their work.

Step 1. The CNs as mentioned above are defined using the work of Garg et al. [8]. Customers are required to rate the performance of the given cloud product in respect to the following criteria: accountability (CN1), capacity (CN2), elasticity (CN3), availability (CN4), service stability (CN5), serviceability (CN6), on-going cost (CN7), service response time (CN8) and security (CN9).

Next, the PTRs that will be used to satisfy the CNs are determined again based on the work of Garg et al. [8]:accountability performance (PTR1), CPU capacity (PTR2), memory capacity (PTR3), disk (PTR4), time (PTR5), availability (PTR6), upload time (PTR7), .CPU stability (PTR8), memory stability (PTR9), free support (PTR10), type of support (PTR11), virtual machine cost (PTR12), inbound data cost (PTR13), outbound data cost (PTR14), storage cost (PTR15), service response time range (PTR16), service response time average value (PTR17) and security performance (PTR18).

Step 2. After having defined CNs and PTRs in the first step, next step involves determining the relative importance of the CNs by asking the following questions: 'Which CN should be emphasized more in establishing the best cloud product?'. We used the same weights for CNs as obtained by Grag et al., as they made the calculations in this step with AHP. AHP and our proposed ANP methodology use the same calculations for this step. The resulting importance weights are given as:

$$\mathbf{W_1} = (0.05 \quad 0.06 \quad 0.04 \quad 0.14 \quad 0.04 \quad 0.02 \quad 0.3 \quad 0.3 \quad 0.05)^T$$

Step 3. Assuming that PTRs are independent, they are compared with respect to each CN, which results in the column eigenvectors regarding each CN. For instance, one of the possible questions can be: 'What is the relative importance of "CPU capacity" when compared to "memory capacity" on controlling "capacity"?'; yielding to the weights presented in Table 1. Similarly, the degree of relative importance of PTRs for the remaining CNs are calculated and presented in Table 2.

Table 1. Relative importance weights of the PTRs for "capacity"

Capacity (CN2)	Relative importance weights
PTR2	0.5
PTR3	0.3
PTR4	0.2

Table 2. The column eigenvectors with respect to each CN

W_2	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	CN9
PTR1	1	0	0	0	0	0	0	0	0
PTR2	0	0.5	0	0	0	0	0	0	0
PTR3	0	0.3	0	0	0	0	0	0	0
PTR4	0	0.2	0	0	0	0	0	0	0
PTR5	0	0	1	0	0	0	0	0	0
PTR6	0	0	0	1	0	0	0	0	0
PTR7	0	0	0	0	0.3	0	0	0	0
PTR8	0	0	0	0	0.4	0	0	0	0
PTR9	0	0	0	0	0.3	0	0	0	0
PTR10	0	0	0	0	0	0.7	0	0	0
PTR11	0	0	0	0	0	0.3	0	0	0
PTR12	0	0	0	0	0	0	0.6	0	0
PTR13	0	0	0	0	0	0	0.1	0	0
PTR14	0	0	0	0	0	0	0.1	0	0
PTR15	0	0	0	0	0	0	0.2	0	0
PTR16	0	0	0	0	0	0	0	0.5	0
PTR17	0	0	0	0	0	0	0	0.5	0
PTR18	0	0	0	0	0	0	0	0	1

Step 4. In this step, we have identified the interdependence among the customer needs by considering each CNs effect on others by using pairwise comparisons. The resulting vectors are summarized in Table 3.

Table 3. The interdependence matrix of CNs

W_3	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	CN9
CN1	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
CN2	0,00	0,65	0,14	0,00	0,00	0,00	0,00	0,00	0,00
CN3	0,00	0,12	0,57	0,00	0,00	0,00	0,00	0,00	0,00
CN4	0,00	0,00	0,00	0,45	0,00	0,00	0,00	0,00	0,00
CN5	0,00	0,00	0,00	0,00	0,56	0,00	0,00	0,33	0,00
CN6	0,00	0,00	0,00	0,29	0,32	0,67	0,00	0,10	0,00
CN7	0,00	0,23	0,29	0,00	0,00	0,29	1,00	0,00	0,00
CN8	0,00	0,00	0,00	0,14	0,00	0,00	0,00	0,57	0,00
CN9	0,00	0,00	0,00	0,12	0,12	0,00	0,00	0,00	1,00

Step 5. Similar to step 4, in this step we have determined the dependence among the PTRs with respect to CNs. The resulting eigenvector of all the pairwise comparisons among PTRs are build using similar pairwise comparisons. Due to space limitations the resulting matrix is not given.

Step 6. In this step, we obtain overall priorities of the PTRs. First, we obtain the interdependence priorities of the customer needs by multiplying the weights obtained in previous steps. Overall priorities of the PTRs ($\mathbf{W_{ANP}}$) are obtained by multiplying $\mathbf{W_A}$ and $\mathbf{W_C}$. The resulting weights are given in Table 4.

Table 4. Overall priorities of the PTRs

	PTR1	PTR2	PTR3	PTR4	PTR5	PTR6	PTR7	PTR8	PTR9
W_{ANP}	0.073	0.084	0.045	0.034	0.028	0.041	0.026	0.064	0.050
	PTR10	**PTR11**	**PTR12**	**PTR13**	**PTR14**	**PTR15**	**PTR16**	**PTR17**	**PTR18**
	0.043	0.022	0.161	0.033	0.037	0.050	0.093	0.062	0.054

The ANP analysis results indicate that the most important cloud service attribute is "VM cost", followed by, "Service response time range" and "CPU capacity".

Step 7. The performance values of each cloud service provider for each PTR are evaluated in this step. The performance values for each PTR are based on the data provided in the work of Garg et al. [8]. The incommensurability issue faced when different units are used to measure the performance is resolved using a normalization scheme. Based on the obtained final ratings for cloud products, the cloud services are ranked as $Provider_1 \succ Provider_3 \succ Provider_2$ with performance values of $\{0.782, 0.761, 0.740\}$. When we compare this result with the work of Garg et al. [8], they ranked the service providers as $Provider_3 \succ Provider_1 \succ Provider_2$. The main reason for the difference is that AHP methodology used in [8] assumes that there is no interdependence among customer needs and no interdependence among product technical requirements. We believe that incorporating dependence issues into the analysis enables to analyze such a complex decision problem in a more complete manner.

6 Conclusion

Cloud services have heterogeneous technical and managerial specifications. Therefore, it is a challenging task to determine which product is better than another. It all depends on the requirements and expectations of the customer. In this paper we proposed that QFD, which simply intends to analyze customers' needs and transform this subjective information into measurable product attributes could be used to identify the most capable cloud service. QFD was chosen as the decision support tool, as it provides a systematic way to combine different sources of data, both subjective like customer expectations, and also objective like product attributes and competitive analysis results. Possible extensions of this work could implement budget and technical constraints to the decision framework, which could have direct influence on the selection process. Going further, the list of CNs and also PTRs could be enriched based on Cloud Service Measurement Index Consortium's measurement indexes [22] and also ISO/IEC 25010:2011 standard.

7 Acknowledgment

This research has been financially supported by Galatasaray University Research Fund, with the project number 13.402.006.

8 References

[1] Mell, P. and Grance, T., 2011, The NIST Definition of Cloud Computing. National Institute of Standards and Technology.

[2] Hoefer, C. N., & Karagiannis, G., 2011, Taxonomy of cloud computing services. Journal of Internet Services and Applications, 2(2), 81-94.

[3] Buyya et al., 2009, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," Future Generation Computer Systems, vol. 25, no. 6, 599–616.

[4] Saaty, T.L., 1996, Decision Making with Dependence and Feedback: The Analytic Network Process, RWS Publications, Pittsburgh.

[5] Saaty, T.L., The Analytic Hierarchy Process, 1980, McGraw-Hill, New York.

[6] Saaty, R. W., 2003, Decision making in complex environments. Pittsburgh: Creative Decisions Foundation.

[7] Carnevalli, J.A., Miguel, P.C., 2008, "Review, analysis and classification of the literature on QFD – Types of research, difficulties and benefits", International J. Production Economics, 114, 737-754.

[8] Garg, S.K. Versteeg, S. & Buyya, R., 2013, A framework for ranking of cloud computing services. Future Generation Computer Systems, 29, 1012-1023.

[9] Qu, L., Wang, Y., Orgun, M.A., 2013, Cloud Service Selection Based on the Aggregation of User Feedback and Quantitative Performance Assessment, Proceeding SCC '13 Proceedings of the 2013 IEEE International Conference on Services Computing, 152-159.

[10] Godse, M., Mulik, S., 2009, An approach for selecting Softwareas a-Service (SaaS) product. IEEE International Conference on Cloud Computing, 155–158.

[11] Garg, S. K., Versteeg, S., Buyya, R., 2011, SMICloud: A framework for comparing and ranking cloud services. In Utility and Cloud Computing, 210–218.

[12] Lo, C., Chen, D., Tsai, C., Chao, K., 2010, Service selection based on fuzzy TOPSIS method, 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, 367-372.

[13] Menzel, M., Ranjan, R., 2012, CloudGenius: Decision Support for Web Server Cloud Migration, Proceeding WWW '12 Proceedings of the 21st international conference on World Wide Web, 979-988.

[14] Ergu, D., Peng, Y., 2013, A framework for SaaS software packages evaluationand selection with virtual team and BOCR of analytic network process, The Journal of Supercomputing, September 2013, DOI 10.1007/s11227-013-0995-7.

[15] Chan, L.K., Wu, M.L., 2005, "A systematic approach to quality function deployment with a full illustrative example", Omega, 33, 119-139.

[16] Saaty, T.L., 1980, The Analytic Hierarchy Process, McGraw-Hill, New York.

[17] Cheng, E.W.L, Li, H., Yu, L., 2004, The Analytic Network Process (ANP) Approach to Location Selection: A Shopping Mall Illustration, Construction Innovation, 5, 83 – 97.

[18] Saaty, T.L., 1996, Decision Making with Dependence and Feedback: The Analytic Network Process, RWS Publications, Pittsburgh.

[19] Saaty, T.L., Vargas, L.G., 1998, Diagnosis with dependent symptoms: Bayes theorem and the analytic hierarchy process, Operations Research, 46(4), 491-502.

[20] Saaty, T.L., Takizawa, M., 1986, Dependence and independence: From linear hierarchies to nonlinear networks, European Journal of Operational Research, 26, 229-237.

[21] Karsak, E.E., Sozer, S., Alptekin, S.E., 2003, Product planning in quality function deployment using a combined analytic network process and goal programming approach, Computers & Industrial Engineering, 44, 171-190.

[22] C. S. M. I. C. (CSMIC), "SMI Framework," URL http://betawww.cloudcommons.com/servicemeasurementindex.

TOWARDS A KNOWLEDGE TRANSFER MEASUREMENT FOR SOFTWARE REQUIREMENTS

José Jairo CAMACHO
Universidad Nacional de Colombia
Bogotá – Colombia
jjcamachov@unal.edu.co

Jenny Marcela SANCHEZ-TORRES
Universidad Nacional de Colombia
Bogotá – Colombia
jmsanchezt@una.edu.co

Ernesto GALVIS-LISTA
Universidad del Magdalena
Santa Marta – Colombia
egalvis@unimagdalena.edu.co

There has been an increasing interest about knowledge transfer (KT) in software engineering last years, but, less in software requirements (SR) and even less in KT measurement. The purpose of this paper is to make an approach to KT measurement for SR. A mapping from the KT process steps against the software requirements process steps was made, looking for a customized KT process according SR particularities then an approach to metrics were defined for each step. Classic SR metrics are both quantitative and qualitative, none of them related directly with knowledge transfer but with knowledge codification and knowledge sharing. This paper presents the SR process as a KT process obtaining KT oriented FACTORS in one approach to KT measurement in SR.

Keywords-component; Knowledge Transfer Process; Knowledge Management; Software Engineering; Software Requirements; Software Metrics

I. INTRODUCTION

Software Engineering has been recognized as a knowledge intensive application discipline(Rus & Lindvall 2002),(Dingsøyr et al. 2009) and (Ward & Aurum 2004). For this reason, in the last decade there has been an increasing interest about knowledge management in software engineering. In particular, the processes of knowledge codification and knowledge sharing have received most attention and they have been researched in diverse ways.

Research done about Knowledge Transfer, KT, in software engineering had been more related to the handling of software knowledge along and among software development organizations, focusing in factors affecting knowledge transfer and the levels of the transfer inside software organizations (i.e. KT between multinationals, projects, teams and people).

Even when KT in software engineering has been studied, there is a lack of research about KT throughout the detailed sub process of software engineering process, which could be seen according SWEBOOK as: software requirements, SR, software development, testing and maintenance. Since SR is the first sub process at the beginning of software projects, the goal of this paper is to describe how the KT happens in SR and it is done mapping KT and SR elements in a matrix. SR is represented as software elicitation, analysis, specification and validation from the SWEBOOK point of view which was elected due to their effort to summarize what is known about the software process. Finally, we want to gain insights on how KT process could be measured, so indicators are proposed for each mapped process step. More specifically, the research questions for this study are:

1. How does the KT take place in SR?
2. Which are factors affecting KT in each stage of SR?

In developing this paper, we start with a background about KT in section II. Next in section III the mapping method is described. Section IV presents the results of the mapping and metrics proposed so research question are resolved. And finally, in section VI the conclusions of this study are presented.

II. ATHEORETICAL BACKGROUND ON KNOWLEDGE TRANSFER

A. Knowledge Transfer concept

On the one hand, knowledge has been defined as the information and experience grouped usefully in some context(Alavi & E. Leidner 2001), and literature shows a consensus about the taxonomy which represents knowledge as tacit and explicit(Nonaka 2007). On the other hand, transfer means to pass an element form one side to another(Watson & Hewett 2006) and(Borgatti & Cross 2003), so, knowledge

transfer means to pass useful information and experience from one context (project) to another (inside or outside of an organization).

Nevertheless, such transfer, according to some authors, cannot be done (Krogh 2003) due to the fact that knowledge is personal and unique. Every time knowledge passes from tacit to explicit, new knowledge is generated so it is different from the previous one(Garavelli et al. 2002). In this way, the exactly KT cannot be possible.

It should be noted that KT is different from knowledge sharing(Kumar & Ganesh 2009)(Argote & Ingram 2000), since the fact that a person shares knowledge does not mean that he/she already did a transfer. Consequently, entity A (person, business unit or company) transfers knowledge to entity B, just when B is able to apply it in a useful way in its own context. By the same token, it can be said that only sharing knowledge has occurred.

Knowledge sharing is important as a KT enabler, but sharing alone is not enough to make transfer occurs. This is remarkable because, until now, the greatest advances in knowledge management applied to software engineering have been done at the level of knowledge sharing using knowledge codification (Garavelli et al. 2002),(Gosain 2007),(Farenhorst & Vliet 2009) and(Souza et al. 2010).

KT is more than mere codification because it demands more than building "knowledge" bases (data and information)(Kumar & Ganesh 2009). Those bases ended being only data repositories because those bases are used just to code the knowledge, however KT is related to a human process and it could only be generated through cognitive process inside people's mind(Carayannis 1999).

B. Knowledge Transfer as a Process of Knowledge Management

KT is one of the most important processes for knowledge management(Kuhn & Abecker 1997),their activities are mainly three, gather the knowledge from a source, code it through a channel, and pass it to a receipt (Albino et al. 2004). KT inside the knowledge management could be seen as a final process, because after create, store and share the knowledge, only when transfer occurs knowledge management makes sense and could be said that is useful(Argote & Ingram 2000) and (Kumar & Ganesh 2009).

KT process could be depicted as a source of knowledge who has explicit or tacit knowledge and a receipt who has to interpret the knowledge so it is able to apply knowledge transferred. It is important to note that for transfer success, Knowledge codification at the source must to be done, because the knowledge at the source, even if explicit, has to be codified in an object with significance for both source and receipt.

C. Knowledge Transfer Barriers and Enablers

KT has barriers and enablers inside an organization, such barriers are related to economic, cultural and social capital(Liebowitz & Suen 2000). Depending on the type of organization some factors are more relevant, for instance in multinational environments, culture differences are more important than in small and medium environments(Ambos & Ambos 2009), (Gera 2012) and (Tichá & Havlíček 2007).

The barriers/enablers factors influencing KT usually come from hypothesis about people behavior, is used to presume that for KT occurs, a predisposition of the people to share knowledge should exist(Karlsen et al. 2011).Typically, factors always have two components, one organizational and one technologic. The organizational is related to the behavior of source and receipt of knowledge at different levels (multinationals, projects, teams and people), some factors include: communication mechanisms, trustiness, commitment, reciprocation, identity, shared goals, culture distance, language distance and geographic distance (Ambos & Ambos 2009; Duan et al. 2010; Garavelli et al. 2002),(Chen et al. 2008) and(Aurum A. 2008). The technological factors are pertaining to tools helping KT process at levels mentioned above, some tools like knowledge bases and ontologies are mentioned as cultural distance helpers, while some others like VoIP help

communication (Chen & Lovvorn 2011) and (Ambos & Ambos 2009).

III. METHOD

KT process stages were mapped against SR process steps. On the one hand the reference for the KT stages was taken from relevant authors according a literature review about KT in SR and their measurement, on the other hand the SWEBOOK was taken into account, for the SR process, because it is an effort from the computer society to characterize what is known about software engineering process, and promote a consistent view of SR. Finally, factors affecting KT were organized for each step mapped.

The literature review was defined as follows: ***step 1***, a set of papers were taken from the result of a search equation in SCOPUS, over title, abstract and key words, the search equation used is as follows: ("Knowledge transfer" or "knowledge sharing") and ("Software requirements" or "requirements engineering") and ("metric" or "measurement" or "indicator"); ***step2***, a quick review over the results of step1 was performed, the review include the reading of abstract, introduction and conclusions, the criteria used for the selection was the pertinence about SR and KT measurement; ***step3***, a full reading over the papers resulting form step2 was done, from this reading a group of KT stages were defined and each KT stage was linked against SR steps, from SWEBOOK, in order to find which KT stage and SR step matches better, the linking was done according common functions and goals from KT-SR; ***step4***, a group of factors affecting the mapping were defined, according authors from step3 papers.

IV. RESULTS

After being applied the methodology explained above, the results found are as follows: ***step1***, resulted in 373 papers; ***step2***, resulted in 49 papers; ***setp3***, resulted in four dimensions where (Szulanski 2000), (Schwartz 2007), (Minbaeva 2007), (Goh et al. 2008) and (Simonin 2004) are the principal references, step3 is detailed in section A. Responding the questions proposed in the introduction, the KT and SR process are mapped, thus, making visible how KT takes place in SR process, and answering first question in the introduction. The ***step4*** resulted in 7 factor groups that are depicted in section B, answering the question two proposed in the introduction.

A. The knowledge transfer process inside Software Requirements process.

Starting with (Szulanski 1996) who states that KT has four stages: Initiation, implementation, "ramp-up" and integration, other authors start using the word KT process adding or modifying steps like: information acquisition, documentation, transmission, source and receiver perception (Verkasalo & Lappalainen 1998), gather the knowledge from a source, code it through a channel, and pass it to a receipt (Albino et al. 2004), Idea creation, sharing, evaluation, dissemination and adoption (Levine & Gilbert 1998).

SWEBOOK divide SR in seven topics: SR fundamentals, Requirements process, elicitation, analysis, specification and validation, Practical considerations and SR tools. But, only four are going to be considered which are the related with the strictly SR process: elicitation, analysis, specification and validation.

In short, there are four dimensions for the knowledge transfer to occur.

1). Initiation: where the decision to KT and information acquisition is done by gathering the knowledge from a source, in a software context it is supposed that the source is motivated enough to share their knowledge because the source is the client who need the software, at this point the KT for software requirements differ from classical KT in organizations, because the receiver of the knowledge (i.e. software analysts) doesn´t intend to apply such knowledge but to build a software specification. This first step match with software elicitation stage for SR, because is where the first approach to business knowledge stake holders is made, those stake holders initiate the sharing of their knowledge and KT starts.

2) Implementation: is about the formal flow of knowledge from the source to the receipt, first software specification which could be seen as the source of knowledge codification occurs, the elicitation step ends and start the analysis of such first requirements, implementation cease or diminish with the software specification because is where the receipt starts using the transferred knowledge (requirements). At this point implementation step for KT differ from classical KT in organizations, because the receipt isn't going to use the knowledge in his behalf, but for a software specification analysis.

3) "Ramp-up": in this step initial knowledge codification and knowledge dissemination ends, software requirements are fully analyzed giving as a result the formal initiation of a software specification document. Consistence and conjecture of requirements are being evaluated. The software specification serve as a basis for agreement between customers and contractors on what the software product is to do as well and what it is not expected to do.

4) Integration: begins after the receipt achieves satisfactory results with the transferred knowledge, the knowledge is adopted and the perception of source and receiver happens. In the software requirements context is about the software specification end, the awareness of needs and ambiguity are evaluated, starting and ending the validation stage of software requirements, resulting in the final software specification.

Table 1.KT in SR will show the mapping, where KT steps appears in the first column and SR stages take place in the first row, E for elicitation, A for analysis, S for specification and V for validation.

Table 1.KT in SR. Dimensions of KT vs steps of SR.

	E	A	S	V
Initialization. *Information acquisition. *Knowledge gathering. *Knowledge sharing. *Knowledge dissemination.	X			
Implementation. *Documentation. *Knowledge codification.	X	X		
"Ramp-up" *Receipt perception. *Knowledge codification. *Knowledge dissemination.		X	X	
Integration. *Knowledge adoption.			X	X

B. Factors affecting each KT and SR mapping.

Based on (Szulanski 2000), (Schwartz 2007), (Minbaeva 2007), (Goh et al. 2008) and (Simonin 2004) work, the next factors are defined for each mapping.

Factors F1, the Initialization and Elicitation are affected by the willingness to initiate transfer and propensity to share which are related with: acknowledgement and attribution, disseminative capacity, interpersonal connection and motivation of the source.

Factors F2, the Implementation and Elicitation are affected by the ease of transfer which is related with stickiness at initiation, stickiness at implementation, motivation, the awareness of need, the ability to transfer, the ambiguity of knowledge, the retentive capacity and modifiability of requirements.

Factors F3, the Implementation and Analysis are affected by the available time/access of source and receipt, reliability of the source, motivation of the receipt, ambiguity of knowledge, awareness of availability, absorptive capacity of receipt, understandability of requirements and its verifiability.

Factors F4, the Ramp-up and Analysis are affected by the requirements degree of conjecture.

Factors F5, the Ramp-up and Specification are affected by requirements internal consistency.

Factors F6, the Integration and Specification are affected by the available time/access of the receipt and source, the awareness of need from the source and the ambiguity of knowledge.

Factors F7, the Integration and Validation are affected by the correctness and completeness of the specification.

V. CONCLUSIONS

A view of SR as a KT process was done using a mapping between KT stages, here called dimensions of analysis, and SR steps, here called aspects.

The mapping serve as an approximation of KT transfer measurement because shows how SR take place in each KT dimension and which factors influence each match.

Knowledge transfer for software requirements is quite different from classical approach to KT, because transfer of a full body of knowledge is not intended, instead of that, needs who relate with some business knowledge are transferred.

Since KT for SR is a special case, classical metrics used for knowledge transfer should not be applied. For instance, metrics oriented to intellectual capital, number of people trained does not make sense for any SR stage.

Further work will include specific indicators for each factor.

VI. BIBLIOGRAPHY

Alavi, M. & E. Leidner, D., 2001. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly: Management Information Systems*, 25(1), pp.107–136.

Albino, V., Garavelli, A.C. & Gorgoglione, M., 2004. Organization and technology in knowledge transfer. *Benchmarking*, 11(6), pp.584–600. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-23744512384&partnerID=40&md5=d222edacdb209fa33b53cefbff581e32.

Ambos, T.C. & Ambos, B., 2009. The impact of distance on knowledge transfer effectiveness in multinational corporations. *Journal of International Management*, 15(1), pp.1–14. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1075425308001026 [Accessed November 6, 2012].

Argote, L. & Ingram, P., 2000. Knowledge Transfer: A Basis for Competitive Advantage in Firms. *Organizational Behavior and Human Decision Processes*, 82(1), pp.150–169. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0749597800928930 [Accessed November 6, 2012].

Aurum A., D.F.W.J., 2008. Investigating Knowledge Management practices in software development organisations - An Australian experience. *Information and Software Technology*, 50(6), pp.511–533. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-40849140976&partnerID=40&md5=d01522133af390c4c087e6f7faa4e697.

Borgatti, S.P. & Cross, R., 2003. A Relational View of Information Seeking and Learning in Social Networks. *Management Science*, 49(4), pp.432–445.

Carayannis, E.G., 1999. Knowledge transfer through technological hyperlearning in five industries. *Technovation*, 19(3), pp.141–161.

Chen, J.-S. & Lovvorn, A.S., 2011. The speed of knowledge transfer within multinational enterprises: the role of

social capital. *International Journal of Commerce and Management*, 21(1), pp.46–62. Available at: http://www.emeraldinsight.com/10.1108/10569211111111694 [Accessed December 1, 2012].

Chen, Y.-J., Chen, Y.-M. & Chu, H.-C., 2008. Enabling collaborative product design through distributed engineering knowledge management. *Computers in Industry*, 59(4), pp.395–409. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0166361507001583 [Accessed December 1, 2012].

Dingsøyr, T., Bjørnson, F.O. & Shull, F., 2009. What Do We Know about Knowledge Management? Practical Implications for Software Engineering. *IEEE Software*, 26(3), pp.100–103.

Duan, Y., Nie, W. & Coakes, E., 2010. Identifying key factors affecting transnational knowledge transfer. *Information & Management*, 47(7-8), pp.356–363. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0378720610000698 [Accessed November 6, 2012].

Farenhorst, R. & Vliet, H. Van, 2009. Understanding How to Support Architects in Sharing Knowledge. In *SHARK'09*. pp. 17–24.

Garavelli, C., Gorgoglione, M. & Scozzi, B., 2002. Managing knowledge transfer by knowledge technologies. *Technovation*, 22, pp.269–279.

Gera, R., 2012. Bridging the gap in knowledge transfer between academia and practitioners. *International Journal of Educational Management*, 26(3), pp.252–273. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84858858899&partnerID=40&md5=76f9dd205f4423f84ce36c1d202bc14a.

Goh, D.H.-L. et al., 2008. Knowledge access, creation and transfer in e-government portals. *Online Information Review*, 32(3), pp.348–369. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-46249087586&partnerID=40&md5=6e4055f8071d4d2f7ddf55131d42d08e.

Gosain, S., 2007. Mobilizing software expertise in personal knowledge exchanges☆. *The Journal of Strategic Information Systems*, 16(3), pp.254–277. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0963868707000066 [Accessed August 29, 2011].

Karlsen, J.T., Hagman, L. & Pedersen, T., 2011. Intra-project transfer of knowledge in information systems development firms. *Journal of Systems and Information Technology*, 13(1), pp.66–80. Available at: http://www.emeraldinsight.com/10.1108/13287261111118359 [Accessed December 1, 2012].

Krogh, G. Von, 2003. Understanding the problem of knowledge sharing. *International Journal of Information Technology and Management*, 2(3), pp.173–183.

Kuhn, O. & Abecker, A., 1997. Corporate Memories for Knowledge Management in Industrial Practice : Prospects and Challenges 2 Knowledge Management and Corporate Memories. *Journal of Universal*, 3(8), pp.929–954.

Kumar, J.A. & Ganesh, L.S., 2009. Research on knowledge transfer in organizations: a morphology. *Journal of Knowledge Management*, 13(4), pp.161–174. Available at: http://www.emeraldinsight.com/10.1108/13673270910971905 [Accessed November 6, 2012].

Levine, D.I. & Gilbert, A., 1998. Knowledge Transfer: Managerial Practices Underlying One Piece of the Learning Organization.

Liebowitz, J. & Suen, C.Y., 2000. Developing knowledge management metrics for measuring intellectual capital. *Journal of Intellectual Capital*, 1(1).

Minbaeva, D.B., 2007. Knowledge transfer in multinational corporations. *Management International Review*, 47(4), pp.567–593. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-34548201654&partnerID=40&md5=cd3b011dfd9b6b752aeb163d4a33e0d2.

Nonaka, I., 2007. The Knowledge-Creating Company. *Harvard Business Review*, 85(August), pp.162–194.

Rus, I. & Lindvall, M., 2002. Knowledge management in software engineering. *IEEE Software*, 19(3), pp.26–38. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0036575072&partnerID=40&md5=0f0cd4fee3689017f97789c6acfc44fe.

Schwartz, D.G., 2007. Integrating knowledge transfer and computer-mediated communication: Categorizing barriers and possible responses. *Knowledge Management Research and Practice*, 5(4), pp.249–259. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-36148975356&partnerID=40&md5=00d72036f5bdbce05602384a116875d1.

Simonin, B.L.. b c d e f g, 2004. An empirical investigation of the process of knowledge transfer in international strategic alliances. *Journal of International Business Studies*, 35(5), pp.407–427. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-7444254131&partnerID=40&md5=092408754467ce9144d1bbc0b9e1f7da.

Souza, Y.L. et al., 2010. A contribuição do compartilhamento do conhecimento para o gerenciamento de riscos em projetos: um estudo na indústria de software. *JISTEM Journal of Information Systems and Technology Management*, 7(1), pp.185–206. Available at: http://www.jistem.fea.usp.br/index.php/jistem/article/view/10.4301%2FS1807-17752010000100008 [Accessed December 1, 2012].

Szulanski, G., 1996. Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal*, 17(SUPPL. WINTER), pp.27–43. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0041751098&partnerID=40&md5=311433b07b40d218693f39a5fe3897dd.

Szulanski, G., 2000. The Process of Knowledge Transfer: A Diachronic Analysis of Stickiness. *Organizational Behavior and Human Decision Processes*, 82(1), pp.9–27. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0001311775&partnerID=40&md5=f56a49022a9c01291b88a446d0f00c4a.

Tichá, I. & Havlíček, J., 2007. Knowledge transfer. *Agricultural Economics*, 53(12), pp.539–544. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-38149096373&partnerID=40&md5=a352080625957094a8a798045b20d69a.

Verkasalo, M.. c & Lappalainen, P.., 1998. A method of measuring the efficiency of the knowledge utilization process. *IEEE Transactions on Engineering Management*, 45(4), pp.414–423. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0032204893&partnerID=40&md5=1616e85f412b0202ef9f51a8167a6673.

Ward, J. & Aurum, A., 2004. Knowledge management in software engineering - describing the process. *2004 Australian Software Engineering Conference. Proceedings.*, (c), pp.137–146. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1290466.

Watson, S. & Hewett, K., 2006. A multi-theoretical model of knowledge transfer in organizations: Determinants of knowledge contribution and knowledge reuse. *Journal of Management Studies*, 43(2), pp.141–173. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-33645135147&partnerID=40&md5=104a7def37f1ed263758ca09937d22b4.

SESSION

KNOWLEDGE AND INFORMATION VISUALIZATION + IMAGING SCIENCE AND APPLICATIONS

Chair(s)

TBA

Efficient Image Segmentation Algorithm for Mobile Devices

Mark Smith
University of Central Arkansas
Conway, Arkansas 72035

Abstract

An efficient image segmentation algorithm utilized for mobile applications running on the iPhone's iOS platform is presented. Mobile devices such as the iPhone have limited CPU and memory resources, thus presenting a more challenging task when implementing complex algorithms such as image segmentation. The image segmentation utilized in this work splits the image into real-world objects that are numbered for the user to either select for further processing. First, a color quantization algorithm is applied to the entire image thus simplifying the image to only 16 available colors. Next, a fast texture measurement utilizing the co-occurrence matrix is applied to entire image using a pre-selected neighborhood of interest. Multiple regions are then automatically merged based on a color comparison measurement extracted at each object's boundary. The resulting regions are then displayed to the user for further analysis or selection. The primary usage of this algorithm is within other mobile apps that require the segmentation of images into realistic objects. Examples of these apps would be those that read bar codes, QVC codes, or OCR text regions. Results are shown for numerous standard image samples and compared with other image segmentation algorithms.

.

1. Introduction

The interest in mobile devices has exploded in recent years, especially the usage of the Apple's iPhone and iPad. These devices allow users to capture pictures and live video instantaneously while processing these images in real-time by an App. The App described in this paper is used for segmenting the image into real-world objects that can be used for further processing. In other words, this App could provide a foundation for other Apps that require an image (or even a video) segmented into realistic objects for further analysis. The image segmentation algorithm described in this work very efficiently processes the image thus minimizing the limited CPU and memory requirements. A novel image segmentation algorithm implemented in this app consists of the following steps:

- Color Quantization to 16 colors
- Fast Texture measurement extracted from the Co-Occurrence Matrix
- Adjacent regions are merged based on a dominant color matching. Algorithm.
- The resulting segmentation provides a realistic set of objects

The following sections outline each of these steps found in the process discussed in this work as well as a results section comparing this algorithm with other comparable systems.

2. Color Quantization and Color Matching

There are tens of thousands of unique colors in a given image and perhaps millions of unique colors across several pictures of a video sequence. The quantization of all possible colors to only a few levels is an important simplification step, since the comparison so many different color possibilities prove difficult when identifying the optimal foundation color to be applied to a region. The image undergoes a standard k-means clustering algorithm [9,17] and 16 quantized colors are extracted from this initial object. The motivation behind using 16 colors is because it has been found that most realistic regions can be represented by this many discrete labels - thus shading, textured regions, etc can be modeled most accurately this way. Before clustering, the original RGB pixel colors are converted to the CIE- L*a*b* color space which has been shown to be perceptually uniform and therefore preserve more accurate distances than the RGB color space, thus providing superior results [14]. The clustering results on the CIE-L*a*b* colors are then converted back to the RGB colors, the main feature used in this system.

The quantized colors in the regions are then compared with the actual colors in the other regions. The colors will be classified in one of two ways:

1. An existing color found in the largest possible region.
2. A new color not found in the region.

The symbol *pcn* will be used to represent the actual color in an additional region whereas *pcp* represents the corresponding matching color in the largest region. A new color is identified in the additional regions by (1) as

$$\| \mu\text{-}pcn\| > \max\|\mu\text{-}pci\| + \alpha\sigma \quad (1)$$

where μ is the mean of the cluster that *pcp* belongs to, pci is the *ith* color belonging to this cluster with $i = 1, 2, \ldots N$, and N is the total number of colors grouped with the cluster. σ is the standard deviation of the distances computed between μ and the colors in its cluster and is given by

$$\sigma = \sqrt{\frac{\sum_{i}^{N} \|\mu - pci\|}{(N-1)}} \quad (2)$$

and α is a scaling factor. We have found that α equal to 2 works well for the application considered in this work. This color-matching step is illustrated in Fig. 1.

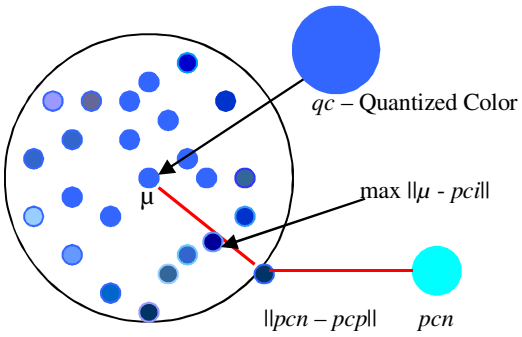


Figure 1. Color Matching

In the example shown in Fig. 1, *pcn* is classified as a new color.

3. Co-Occurrence Texture Measurement

The Gray-Level Co-occurrence Matrix (GLCM) is one of the most popular statistical texture measurements [15,19] and has been used as the primary component in a wide range of image segmentation applications [18,20]. The GLCM is a second-order statistical measurement; second-order statistics take into account the relationship between groups of two (usually neighboring) pixels in the original image. In contrast, first-order statistics, (e.g., mean and variance), do not consider any neighborhood associations. The process by which the GLCM is computed is outlined as follows

1. The GLCM computation utilizes the relation between two pixels at a time; one is called the reference and the other the neighbor pixel.
2. A displacement vector d, as specified as

$$d = < d_x, d_y > \quad (d_x \text{ - distance in horizontal direction})$$
$$(d_y \text{ - distance in vertical direction})$$

is selected and determines the relationship between the pixels in the image. Utilizing only neighboring pixels (d = 1) is the most commonly used distance measurement and is also the one utilized in this system.
3. There are 8 possible relationships (i.e., displacement vectors) that can be formed between neighboring pixels (directions between neighboring pixels are shown in parenthesis –the first component refers to the horizontal displacement, whereas the second parameter refers to the vertical displacement):

< 1,0> (0)
< 0,1> (90)
<-1,0> (180)
< 0,-1> (270)
< 1, 1> (45)
<1,-1> (315)
<-1, 1> (135)
<-1,-1> (225)

4. A displacement vectors d is chosen for each co-occurrence matrix calculation [4,11]. All occurrences of gray levels i and j of two pixels separated by displacement vector d are accumulated. For instance, if $i = 0$, and $j = 0$, and the displacement vector is <1,0>, the calculation is performed by accumulating the frequency on the selected image region that a pixel with gray level 0 (neighbor pixel) falls to the right of another pixel with gray level 0 (reference pixel). The GLCM is a very compact and optimal measurement. An example illustrating a complete GLCM calculation, consider the following 4x4 image with pixel values shown in Fig. 2:

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Fig. 2 Example 4x4 Image Gray-Levels

For example, if the East displacement vector is chosen (i.e., <1,0>), each image pixel is selected in turn as a reference pixel. The pixel immediately to its right is then chosen as the neighbor pixel. The occurrences of these two pixels together are then accumulated. In this example, 0-0 occurs twice, 1-1 occurs twice and so on. The entire co-occurrence matrix for the image in Fig. 2 and the <1,0> displacement vector is shown in Fig. 3.

	0	1	2	3
0	2	2	1	0
1	0	2	0	0
2	0	0	3	1
3	0	0	0	1

Fig. 3 GLCM Computed for Fig. 2

Both the rows/columns pertain to a discrete gray-level 0,1,2, or 3. Note that the co-occurrence matrix is square and its dimensions are always determined by the number of gray-levels (i.e., for this system number of quantized colors) of the image [6,13]. The GLCM dimensions are CxC where C is the largest gray-scale value, or number of quantized colors.

4. Image Segmentation Algorithm

This system uses the mean of the GLCM as a key feature in its image segmentation algorithm (see section 4.) and the mean of the GLCM and its magnitude is given below as [12]:

$$\mu_i = \sum_{i=1}^{C}\sum_{j=1}^{C} i(P_{ij}) \quad (3)$$

$$\mu_j = \sum_{i=1}^{C}\sum_{j=1}^{C} j(P_{ij}) \quad (4)$$

$$\| \mu \| = \sqrt{\mu^2{}_i + \mu_j{}^2} \quad (5)$$

Where μi is the horizontal mean, i is a given row value, Pij is an element of the GLCM, μj is the vertical mean, j is a given column value, ||μ|| is the mean's magnitude, and N is the size of the sliding window used in computing the GLCM mean. The two values, μi and μj are equal because of the GLCM's symmetry [7]. Plotting the magnitude of the GLCM's mean as a 2-d image is shown below in Figure 5:

Fig. 4. GLCM Mean Feature

The mean textured image is very smooth and almost all micro-textures have been removed [8]. The individual region interiors possess consistent gray-scales throughout this image; therefore the region boundaries can be identified from a basic edge detection filter. An NxN filter computing the variance as is selected as the edge detection filter and is convolved with the mean textured image resulting in an edge intensity image. In equation (4) above, μ is the average grayscale within the NxN region and gij is the gray-scale at the ith row, jth column of the GLCM mean. A plot of the edge image generated as a result of applying the above filter is shown below in Figure 5.

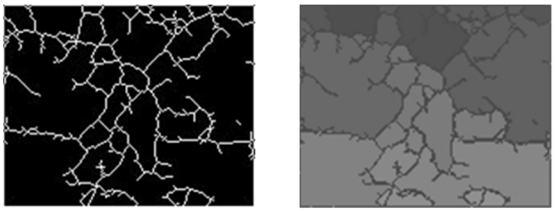


Fig. 5. Image Segmentation Results

The highlighted regions provide an initial set of objects.

5. Region Merging Algorithm

The objects created from the image segmentation of section 3 are occasionally over-segmented thus requiring an additional step to merge similar objects together [10,16]. A step that merges smaller regions with the larger, adjacent region is needed to provide optimal object segmentation. The region merging algorithm introduced in this section demonstrates that small color samples extracted near the boundaries of adjacent regions provide an excellent criteria for merging the areas. The algorithm utilized in this system relies on the dominant (quantized) colors when comparing adjacent regions. Therefore, the adjacent regions are merged based on how similar their colors are to the largest region. The example shown below is for a standard image. The algorithm is summarized as:

1. The regions created by the image segmentation are extracted. The regions (and their corresponding labels) as well as their contours overlaid onto the original color frame are shown in Figure 3.

2. All neighboring segments for each region are determined and only those neighboring segments that are larger are considered as merging candidates. The main concept is that smaller regions are only merged with larger, bordering regions. For example, region 12 has larger

neighboring segments 5,11,16 and 17, whereas region 17 has larger adjacent segment region *5*.

3. Each region's quantized colors are then compared with the quantized colors of each of its larger, neighboring segments. The smaller region will be merged with the larger one if their quantized colors are sufficiently close [5]. The steps utilized in this process are outlined as follows:
   a. A windowed area running the length of the adjacent boundary between neighboring objects is selected for each region. Each area provides a representative sample of the quantized colors for the object. Colors selected at their adjacent boundary provide the best measurement on whether the objects should be merged, thus minimizing the effects from outlying colors. The sampled regions usually have a maximum width of 5 pixels and are parallel to the entire length of the boundary. Additional points are selected when the sampled regions consist of 25 pixels or less. Examples of these sampled regions are shown in Figure 6 for selected neighboring objects.

Fig. 6. Selected Regions

   b. Each quantized color (i.e., discrete label) and its corresponding concentration (measured in percentage) are extracted from each sampled area within each region. Only those quantized colors with a concentration greater than 5% are considered.
   c. If the majority of the quantized colors of the smaller region match those of the larger region, the larger region is then selected as a candidate for merging with the smaller one.
4. Step 3 is repeated for all larger neighboring objects and all candidates for merging with the smaller objects are maintained [2].
5. The candidate which best matches the smaller object's quantized color concentration is then selected as the best matching region for merging. The smaller region is then marked for merging with the larger region – but the actual object merging is not done at this time.
6. Steps 1 – 5 are repeated for all remaining objects.
7. All smaller objects previously marked for merging are then merged with their best matching neighboring objects.

The results of this algorithm as applied to the original segmentation, Figure 3, is shown in Figure 5.

Fig. 7 Region Merging Results

6. Results

The image segmentation algorithm described in this work was implemented as an App on Apple's iPad iOS 7.1 platform using XCode and Objective-C. Testing the App was performed using a series of standard test images added to the iPad's camera roll [9]. The test pictures consisted of popular images extracted from three different categories – *Happy Granny, Foreman*, and *Tennis*. Examples of these test images are shown below:

Fig. 8. Test Images

The author's algorithm was compared with 2 other popular image segmentation algorithm commonly referenced in the literature. The first of these was the JSEG [1] algorithm developed by researchers at UCSB while the second algorithm is implemented as part of the OpenCV [3] library, a popular image processing library implemented by Intel. The algorithms utilize both color and texture when segmenting images. The algorithms were compared based on their speed in milliseconds required to process each of the test images on the iPad device. A table showing the results of these comparisons is shown below in Table I:

Table 1

Algorithm	*Happy Granny*	*Foreman*	*Tennis*
Author's	875	1478	583
JSEG	923	1367	1033
OpenCV	2154	3382	1932

As observed from the table, the results of the author's algorithm appear to be very promising. The algorithm described in this work could provide the substrate layer needed for many apps implemented on the iOS platform requiring a captured image segmented into realistic objects. The applications for this type of algorithm is numerous, especially when segmentation of specific regions such as barcodes or printed text is required.

Future work for this system include enhancing various apps (e.g., barcode, text) with the image segmentation algorithm described in this work thus providing them with the efficient object segmentation capabilities often only currently found in high-end desk-top applications
:

7. 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.

[18] L. S. Davis and S. Johns, "Texture analysis using generalized co-occurrence matrices, " *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol 3, pp. 251-259, 1979.

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

[20] J. Haddon and J. Boyce, "Image segmentation by unifying region and boundary information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, October 1990.

Interactive Visualization of Business Births and Deaths in the U.S. Economy using a Novel Visualization Technique Called *HiFi Pie*

Leonidas Deligiannidis

Professor of Computer Science
Wentworth Institute of Technology
Boston, MA, USA
deligiannidisl@wit.edu

Erik Noyes

Associate Professor of Entrepreneurship
Babson College
Babson Park, MA, USA
enoyes@babson.edu

Abstract

HiFi Pie is a novel technique for interactive information visualization. To illustrate its strength, we explore historic data on new business births and deaths in the U.S. economy. As U.S. Economic Census data continually improves to track the birth and death of new businesses, one can visualize patterns of creative destruction in the U.S. economy, particularly broad historic patterns of whole-economy expansion, all aggregated from industry-level views of new business births and deaths. A novel, interactive viewer built on the visual language the common pie-chart, HiFi Pie allows economic development organizations focused on entrepreneurship, as well as economists and other policy makers, to visualize new business creation trends, including dynamics of entrepreneurial job creation and industry innovation.

Keywords: Information Visualization, Creative Destruction, Entrepreneurship, Economic Development, Job Creation, Visualization Methods

1. Introduction

As data improves to track new business births and deaths in the U.S. economy, new methods are needed to visualize historic patterns of new business creation and industry growth. Economic development organizations focused on entrepreneurship, as well as economists and policy makers, understand that new entrepreneurial ventures are the single greatest driver of new job creation and economic growth, but fast-growth industries are also characterized by high rates of new business failure.

Creative destruction, an idea first advanced by Joseph Schumpeter [1], is the concept that the constant process of new business creation, even driving other businesses to their death, accelerates innovation and economic development. The birth and death of new ventures, while a source of disruption and unemployment, is on the whole an economic win for societies because surviving, competitive ventures create new innovations, economic value and jobs.

These two perspectives, one of whole-economy growth and constituent new venture birth and death are rarely combined to visually provide a coherent sense of their interrelationship.

HiFi Pie represents an effort to explore the potential usability of uniting these two perspectives in a browsable, controllable visual interface. The user-friendly visualization leverages the visual language of the common pie chart, while including added dimensional information about new business births and deaths and industry expansion. Diverging from traditional pie chart conventions, total circle area has precise numerical meaning (i.e., beyond 100%), varies depending on data depicted, and here represents concentric rings of economic and business activity similar to the growth rings of a tree, conveying both percentages and totals.

While the paper details a first, functioning prototype of HiFi Pie, examining nine major industry groups, or sectors, the tool is being developed to visualize multiple levels of industry classification data (e.g., according to the 6-digit, hierarchical, North American Industrial Classification System)

2. Background

The concept of using a pie chart as a visualization mechanism was conceived two centuries ago [2]. William Playfair used the pie chart concept in [2] to visualize the areas, population, and revenues of European countries. In [3] Ian Spence provides a great picture of how the original pie chart evolved over the years and how a pie chart could be used as a useful visualization metaphor.

The authors in [4] provide a wonderful historical review of radial visualization tracking its roots in centuries in the past. One of the concepts that is mentioned in this paper and is related to HiFi Pie is that longer radii of concentric circles have bigger area. The radii, thus, must be computed accurately to reflect the areas added onto the concentric circles. This introduces confusion for the reader and that's the reason the pie-charts do not have widespread acceptance. However, we address this issue by our second stage of our algorithm in which we align the segments each pie slice is composed of.

Krona [5] combines a variant of radial displays with parametric coloring and interactive polar-coordinate zooming. The display resembles pie-charts with inter-segmented pie slices producing an embedded hierarchical visualization. The inter-segmented slices are arranged from the top level of the hierarchy and progress outwards.

In [6], the authors use tree maps with a pie transformation which results in twisted rectangles around a center point to create pie-chart-like visualizations.
In [7] a new framework is proposed for visualizing tables, proportions, and probabilities. They produce pie-charts with different radii for each concentric circle.
PieTrees [8] are area based tree visualization that can be used to represent hierarchical numerical data. They map size directly onto area into a circular layout. Users can expand and collapse any nodes or the entire graph.
Circle View [9] combines pie charts with a novel arrangement of time events on circle segments where each segment is further divided into sub-segments to visualize the distribution and changes over time. This enables it to present a visualization with both local detail and global context in a single view. The resulting visualization graphs resemble pie-charts with different concentric ring widths.

3. The Data

We obtained the data from the US Department of Commerce, United States Census Bureau [10]. Description of the fields in the dataset is located at [11]. The dataset consists of nine sectors:

1. AGR (Agriculture)
2. MIN (Mining)
3. CON (Construction)
4. MAN (Manufacturing)
5. TCU (Transportation)
6. WHO (Wholesale Trade)
7. RET (Retail Trade)
8. FIRE (Finance)
9. SRV (Services)

The dataset contains data from 1977 to 2011. There are many fields in the dataset such as, *firms* (number of firms), *estabs* (number of establishments, *job_creation* (number of jobs created over the last 12 months, etc. The complete list of fields is located at [11].

4. The Algorithm

The algorithm to generate the HiFi Pie consists of two stages. Stage's one goal is to calculate the correct radii of all concentric circles, and thus, the area of each pie slice is represented with high accuracy. The second stage consists of the alignment process where each layer of the pie slice is aligned with all the other layers to produce the final HiFi Pie slices.

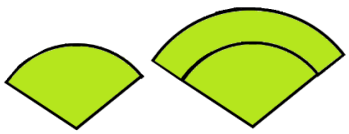


Figure 1. Layers added to each pie-slice, which is data added each year.

Each pie slice consists of layers, data that is added each year and with different rate as shown in figure 1. These layers are centered along their center axis.

4.1. Stage One – Calculating Radii

The algorithm of the first stage consists of several short steps. We present the algorithm in detail (as pseudo-code) and for each step of the algorithm we show how we fill in our data-structures; with the ultimate goal to calculate the radii of each concentric circle to build the pie-slices. We present the data in our data-structures as a table (in the appendix) with the results filled in by the execution of each step of the algorithm. To explain the algorithm with numerical values, we use a small dataset consisting of three years and three sectors (AGR, MIN, and CON). This algorithm needs to be run for each field; in this example we run it for just one field. Initially, the data loaded from our database looks like Table 1.
We first calculate the overall sum of all years of all sectors as shown below; in this example this total sum is equal to 300.

```
for each year Y do
  for each sector S do
    TOTAL += getVal(Y, S)
  done
done
```

We then calculate the percentage of each sector compared to the overall sum calculated in the previous step (also see Table 2 in the appendix).

```
for each year Y do
  for each sector S do
    ValuePercentage = getVal(Y,S) * 100 / TOTAL
    Store(Y,S, ValuePercentage)
  done
done
```

We then compute the sum of the percentages for each year as well as the cumulative sum, as a percentage, starting at the first year in our dataset. In other words, for each year we calculate the percentage sums of all sectors (in the column `TotalForRow`) as well as the cumulative percentages sum of all years (in the column `Cum`) (also see Table 3 in the appendix).

```
Cum = 0
for each year Y do
  V = 0
  for each sector S do
    V = V + getVal(Y,S)
  done
  storeTotal(Y,S,V)
  Cum += V
  storeCum(Y,S,Cum)
done
```

Based on the percentages of each sector, we calculate the degrees/angle of the pie slice that corresponds to each sector for each year (also see Table 4 in the appendix).

```
for each year Y do
  for each sector S do
    Val = get_ValuePerc (Y,S)*360/getRowTotal(Y)
    storeDegrees(Y,S,Val)
  done
done
```

In the final step of stage one of the algorithm, we calculate the corresponding radii for each year in relation to the unit circle. The radius of the unit circle is calculated as `R=sqrt(Area/PI)=sqrt(getCum(year=1979)/PI)` where the `year` is equal to the last – most resent – year in the dataset; all data will be within the circle defined by this radius. The total area of all our data is equal to the cumulative sum which is stored in the `Cum` column for the last year. We use this radius as the radius of the unit circle and we calculate all other radii for all other circles as shown below (also see Table 5 in the appendix):

```
R=sqrt(Area/PI)=sqrt(getCum(year=1979)/PI)
for each year Y do
   Val = sqrt(getCum(Y) / PI) / R
   storeRadius(Y,Val)
done
```

The user, using a slider in the application, can adjust the radii of the circles to better view the pie-charts. We call this user-specified value `sliderVal`. To draw each concentric circle, we simple need to multiply each radii of a corresponding circle with this value:

```
radius = getRadius(Y) * sliderVal
```

where `Y` is the corresponding year.

Now that we have all the data we need, we can contrast the difference in the pie charts we generate using HiFi Pie and using Excel's built-in 2D-piechart Doughnut representation. Figure 2 shows the two generated pie charts. The left one is the data representation using Microsoft Excel's doughnut 2D pie chart. All circles are of the same width. The right one is the HiFi Pie chart before stage two, which is the alignment process of the algorithm. The concentric circles here are shown with different widths. Each radii is correctly calculated to show the correct area each sector occupies, keeping the overall area consistent with the data and the overall area.

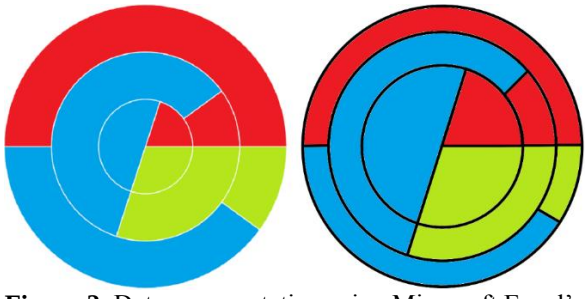


Figure 2. Data representation using Microsoft Excel's Doughnut 2D pie chart graph (left), and using HiFiPie (right) before the alignment process. Notice the width of the concentric circles that are different in size in HiFi Pie, which accurately visualizes the area of each year.

To present the data as an overall conventional pie-chart using Microsoft Excel, we need to calculate the sums of each sector and their corresponding angle of each pie-slice as shown below (also see Table 6 in the appendix):

```
for each sector S do
  Val=0
  for each year Y do
    Val = Val + getVal(Y,S)
  done
  StoreVtotal(S,Val)
  Degrees = (Val * 100 / TOTAL) * 360/100
  storeTotalDegrees(S, Degrees)
done
```

This produces the chart shown in figure 3. Even though the overall data is represented just fine here, what is missing is how each pie slice evolved over the years. HiFi Pie is able to visualize how the data from each year shapes each pie-slice. Stage two of the algorithm will align each layer of each pie slice to produce the final product.

Figure 3. Using a conventional pie-chart to visualize overall data.

4.2. Stage Two – Alignment Process

The results of stage one is fed to stage two of the algorithm. Each year adds a layer to the previous years. When considering all sectors' data for a particular year, this additional data that is added appears as a ring of data. For this example, the three rings of data for the three year are shown in figure 4. We construct the layers by drawing layer N and then subtracting layers N-1 and beyond, recursively, where N in the most current layer representing the data for the most recent year. These layers (or rings) contain the data for a particular year for all sectors. We then extract the segments of the rings of each sector and center/align them producing the HiFi Pie slices as shown in figure 5.

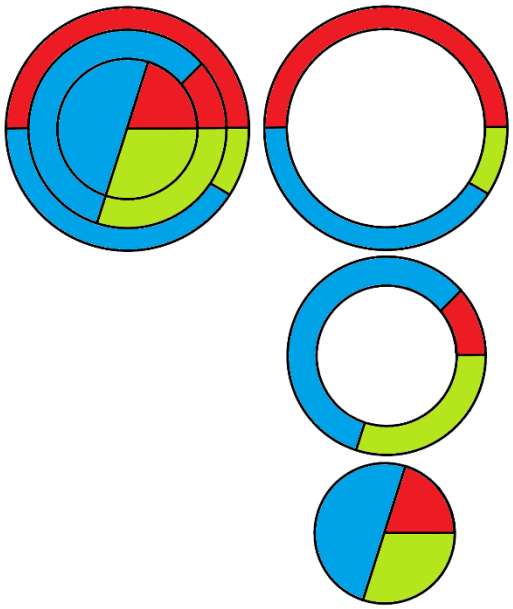


Figure 4. The Layers (rings) of the HiFi Pie. Each layer represents data added to the previous year.

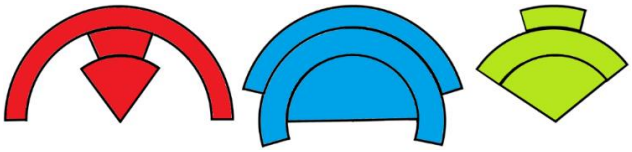


Figure 5. The HiFi Pie slices after the alignment process. Each year for each sector in centered to reveal each HiFi Pie slice.

5. The Graphical Interface

The graphical interface presents to the user the tools to select the different *fields* to be visualized, modify the range of years the HiFi Pies represent, to zoom in and out on a HiFi Pie chart, and to expand the pie slices. Expanding the slices means that we move all slices away from their common center to enable the user to rotate them later and fit them next to each other. It also enables the user to direct-manipulate the entire chart or individual HiFi pie slices; the user can mouse click on a slice to move it and rotate it; left mouse button moves the entire HiFi Pie, shift-left mouse button moves individual HiFi Pie slices, and right mouse button rotates a HiFi Pie slice around the HiFi Pie's common center. From the menu bar the user can choose the quality in which the HiFi Pie slices are drawn; select/deselect antialiasing, and level of detail to render the HiFi Pie slices faster. The user can also select the Universal Radius option so that each HiFi Pie chart is drawn with respect to the largest HiFi Pie. We first calculate the total values of each *field* and we get the maximum. From then on, each HiFi Pie is drawn with respect to this maximum radius. This enables us to compare different HiFi Pie chart side by side. Figure 6 shows the HiFi Pie before we run the stage two of the algorithm. The radii are shown on top of the pie chart and by themselves alone to show that the radii are different for each concentric circle. The area of the outside rings grows faster than the rings' area closer to the center.

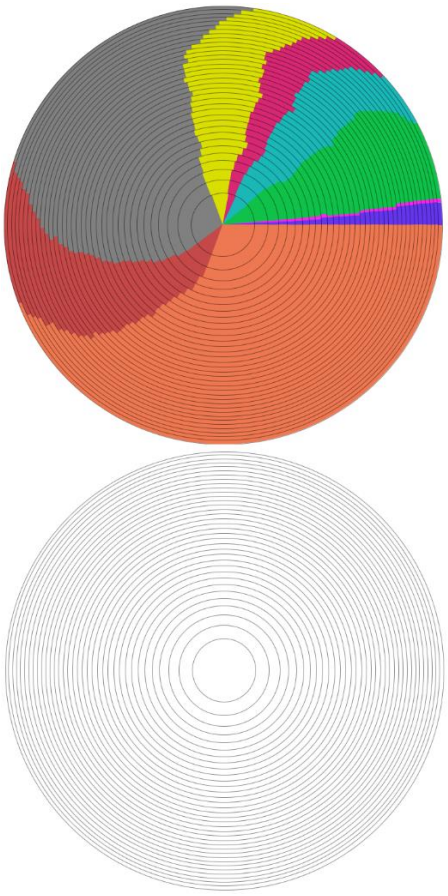


Figure 6. HiFi Pie of the *Estabs* field before stage two of the algorithm. The radii of the concentric circles is shown on top of the data as well as by themselves.

After stage two, where we align and center each sector data from Figure 6, we produce an in-memory image (for each sector) where we render on it the aligned pie slices to produce the HiFi Pie slices as shown in figure 7. Because of the shapes of the HiFi Pies, it is very common, that the individual HiFi Pie slices overlap. The user, however, can expand the char, moving all sectors away from their common center, then rotate and move the individual slices to produce HiFi Pie charts where there is no slice overlaps as shown in figure 8.

The High Fidelity of the HiFi Pie slices can be seen in figure 9. This is a close up view of figure's 8 bottom left area. Here we can see with high precision how the data from individual years affected the formation of each pie slice.

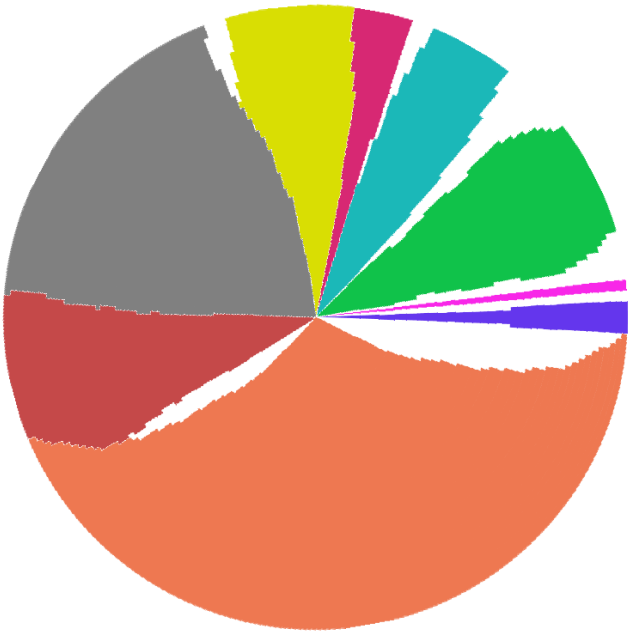


Figure 7. Visualization of the Establishments field for all years using HiFi Pie; the final result.

Figure 8. Result of the HiFi Pie of figure 7 after the user expands the HiFi Pie to avoid slice overlaps.

Figure 9. Close up view of figure's 8 bottom left area where we can clearly see the impact of the data that is added each year.

6. Conclusion

HiFi Pie demonstrates the potential viability and usability of a multidimensional, interrelated representation of creative destruction leveraging the pie chart structure. The visualization captures economic growth, industry growth, and new business births and deaths. Applications of HiFi Pie may include use by policy makers or economic develop organizations focused on entrepreneurship to contextualize policy, as well as use in entrepreneurship educators to depict and explore processes of creative destruction. Further research and testing is needed to explore industry views for multiple levels of analysis (e.g., according to the 6-digit, hierarchical, North American Industry Classification System) to provide finer-grained views of new business creation and job creation beyond broad industry groups such as agriculture, mining and services. One challenge with representing multiple, smaller slices of HiFi Pie is the breakdown of the classic pie chart structure and the potential that relative percentages and areas for different slices become more difficult to compare. However, it is quite possible that the basic pie chart structure is flexible and adaptable to integrate added dimensions of information while retaining core visualization properties of establishing relative areas. One important area for future research is the opportunity to establish meaning in the circular layout of industry categories. For example, the Internet and web industry has most shaped the services industry, suggesting a logic to put most-related industries side by side (potentially based on related products or related customers). While some arrays of basic pie charts have an underlying logic for category proximity (e.g., percentage of respondents who are 18-24 years old adjoining percentage of respondents who are 25-32), there are likely means to automatically array industry segments based on relationships.

Because the study of composition and change is central to many types of analysis and presentation, of course HiFi Pie may be extended to represent and browse other types of hierarchical classification data. Broadly, this research

aims to bring together research on entrepreneurship and economic development with research on interactive information visualization to suggest new opportunities and directions at the intersection of the two specialized fields.

References

[1] Joseph Schumpeter, The Theory of Economic Development, Cambridge, MA: Harvard University Press, 1934.

[2] William Playfair "The Statistical Breviary". Printed by T. Bensley, Bolt Court, Fleet Street. London 1801.

[3] Ian Spence "No Humble Pie: The Origins and Usage of a Statistical Chart" Journal of Educational and Behavioral Statistics Winter 2005, Vol. 30, No. 4, pp. 353–368.

[4] Geoffrey M. Draper, Yarden Livnat, and Richard F. Riesenfeld. "A Survey of Radial Methods for Information Visualization", IEEE Transactions on Visualization and Computer Graphics, Vol. 15, No. 5, Sep/Oct 2009.

[5] Ondov et al.: Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 2011 12:385.

[6] Roel Vliegen, Jarke J. van Wijk, Erik-Jan van der Linden "Visualizing Business Data with Generalized Treemaps", IEEE Trans. On Visualization and Computer Graphics Vol12, No. 5, Sep/Oct. 2006.

[7] Hadley Wickham and Heike Hofmann, "Product Plots", IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12, Dec 2011.

[8] Richard O'Donnell, Alan Dix, Linden J. Ball "Exploring the PieTree for representing numerical hierarchical data". In: People and computers XX — engage. Springer, London, pp. 239-254. ISBN 978-1-84628-588-2, 2007.

[9] Daniel A. Keim, Jorn Schneidewind, Mike Sips. "CircleView A New Approach for Visualizing Timerelated Multidimensional Data Sets". In Proc. Of Advanced Visual Interfaces (AVI), May 25-28, 2004, Gallipoli (LE), Italy.

[10] Home page of the "US Department of Commerce, United States Census Bureau". URL: http://www.census.gov/ces/dataproducts/bds/, retrieved Apr. 2 2014.

[11] Fields description from the "US Department of Commerce, United States Census Bureau". URL http://www.census.gov/ces/pdf/BDS_Codebook.pdf, retrieved Apr. 2 2014.

Appendix

Table 1. Initial load of data in our data-structure.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20			50			30					
1978	10			60			30					
1979	50			40			10					

Table 2. Calculating percentages of each sector for each year.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20	20/3		50	50/3		30	30/3				
1978	10	10/3		60	60/3		30	30/3				
1979	50	50/3		40	40/3		10	10/3				

Table 3. Calculate total percentages for each year as well as their cumulative sum.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20	20/3		50	50/3		30	30/3		100/3	100/3	
1978	10	10/3		60	60/3		30	30/3		100/3	200/3	
1979	50	50/3		40	40/3		10	10/3		100/3	300/3	

Table 4. Calculate the angle of the pie slice for each year for each sector.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20	20/3	72	50	50/3	180	30	30/3	108	100/3	100/3	
1978	10	10/3	36	60	60/3	216	30	30/3	108	100/3	200/3	
1979	50	50/3	180	40	40/3	144	10	10/3	36	100/3	300/3	

Table 5. Calculate the radii of each sector for each year.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20	20/3	72	50	50/3	180	30	30/3	108	100/3	100/3	0.577
1978	10	10/3	36	60	60/3	216	30	30/3	108	100/3	200/3	0.815
1979	50	50/3	180	40	40/3	144	10	10/3	36	100/3	300/3	1

Table 6. Calculating the overall area to display the data using a conventional Pie Chart.

	Sectors											
	AGR			MIN			CON			TotalFor		
Year	Value	%	degrees	Value	%	degrees	Value	%	degrees	Row %	Cum %	Radius
1977	20	20/3	72	50	50/3	180	30	30/3	108	100/3	100/3	0.577
1978	10	10/3	36	60	60/3	216	30	30/3	108	100/3	200/3	0.815
1979	50	50/3	180	40	40/3	144	10	10/3	36	100/3	300/3	1
	80		**96**	**150**		**180**	**70**		**84**			

SESSION

DATABASES, INFORMATION RETRIEVAL AND SEARCH + BOOKMARKING METHODS + AGENT TECHNOLOGIES

Chair(s)

TBA

Bookmarking and Tagging Patterns in Social Bookmarking Systems

Alawya Alawami[1]
School of information Science, University Of Pittsburgh, Pittsburgh, PA,USA

Abstract - *Social bookmarking systems are a promising tool for classifying web resources. To be able to use such systems intelligently, it is important to understand how they work. This study takes a look at one such system to examine tagging and bookmarking patterns. The existing literature tends to view these patterns somewhat simplistically. We specifically examine three questions: (1) at what rate do bookmarks accumulate? (2) To what extent do early tags influence later tagging behavior? (3) How do the top 10 tags evolve? Given the influence of the literature on the direction of future research on social bookmarking systems, a more careful analysis may be of use.*

Keywords: Tagging patterns, Social Bookmarking Systems, Tagging behavior.

1 Introduction

With the development of the social web, many interesting tools that focus on the social interaction and collaboration between users have appeared including Facebook, Delicious, Flicker, CiteULike, Twitter etc. These social systems provide new forms of interaction between users and resources. Social bookmarking systems (SBS) such as Delicious, CiteULike , Dogear, etc. are sometimes defined as "Public link management systems"[1]. They were first created to provide users with bookmark portability. They have shown the ability to attract a significant number of users who have collected a large set of resources. They have also attracted the research community because of their ability to generate a significant amount of content in a short period of time. Unlike library classification systems, social bookmarking systems provide tools to classify documents in an unrestricted way where "classification" arises from uncoordinated agreement among users on terms that describe documents. This classification has been called 'Folksonomy' by Thomas Vander Wal [2]. To optimally use the data generated by these systems it is important to understand the nature of the systems. This research takes a closer look at the nature of such systems by utilizing Delicious, which is one of the early social bookmarking systems with a significant number of public users.

2 Previous Work

Social Bookmarking Systems (SBSs) contain a large number of bookmarks that associate users and resources. Tagging provides a mechanism to both organize resources and facilitate browsing and discovery of new resources. Tags are also being used to supplement indexing and ranking as described in [3-6]

Some research focuses on understanding how folksonomies develop. Other researchers study tagging behavior: how users tag resources, what motivate users toward those systems, what makes some resources more popular than others, how the vocabulary evolves in SBSs, etc.

An early study done by Golder and Huberman [7] analyzed a small set of resources (212 resources) and found many regularities in the system. They found that tagging patterns tend to stabilize after about 100 bookmarks. Furthermore, they found that early tags tend to be more popular. They observed that early tags are more general and that they provide a sense of agreement among users which also implies that later tags patterns where mainly influenced by earlier ones. They relate the stabilization of this pattern to two factors: imitation by other users and shared knowledge. They argue that since Delicious provides users with a list of tag suggestions; initial tags influence later taggers[7].

A similar study done by Wetzker et al. [8] analyzed a data set of 142,341,551 bookmarks. In their work; they looked at delicious system activity as a whole. They found that

- Users activity follows a power law distribution
- Burst in the popularity of resources are caused by their appearance on delicious main page or another popular site or it could be the similar interests of users on the network.
- "Delicious community pays attention to new resource for a very short period of time. As a result URLs receive most of their posts very quickly and disappear shortly afterwards".

Halpin, Robu and Shepherd [9] looked at the dynamics of SBSs and the possibility of defining collaborative systems as complex systems. They found that tagging frequency follows a power law distribution. Marlow et .al. [10] conducted a study that classified systems. They argued that user behavior is influenced by many factors such as: tagging rights, tagging support, aggregation, type of object being tagged, source of material, etc. They also discuss two motivations behind using SBSs: organizational and social.

Farooq et al. [11] looked at set of data from CiteULike and Delicious; in their work, they identified six tags metrics:

1. Tag growth: looking at how the tag vocabulary evolves over time
2. Tag reuse: measuring how many times tags in the system have been reapplied

3. Tag non-obviousness: comparing the tag to its resource to see if the tag helps describe the resource.
4. Tag discrimination: looking at how the tag discriminates a resource
5. Tag frequency: characterizes tags based on their usage frequencies
6. Tag patterns: how users tag resources over time

Some of these studies were done during the early stages of Delicious' history when the system did not have much data[7]. Kipp and Campbell [12] focused on frequency and co-word analysis to investigate whether collaborative tagging could support library cataloging and indexing. Farooq et al. [11] focused their analysis on academic paper social bookmarking which might be applicable to other SBS such as Delicious. Although they have mentioned many useful metrics which we are looking at, they did not examine the patterns that can arise from those metrics. Golder and Huberman [7] looked at different tagging patterns and usage in SBS. They have also analyzed user's behavior but their analysis was limited to two small sets (212 of popular resources with their bookmarks, 229 users with their bookmarks) and their analysis was also done when Delicious was still in its infancy. The study by Wetzker et al. [8] looked at Delicious users and resources as a whole.

In this study we extend their work. We examine some of the findings of previous work by analyzing a set of 41,469,488 resources. With this data set we examine three questions: (1) at what rate do bookmarks accumulate? (2) To what extent do early tags influence later tagging behavior? (3)How do the top 10 tags evolve?

3 The Data Set

Delicious was created by Joshua Schachter in 2003 and acquired by Yahoo in 2005. In 2008, Delicious announced that they had reached 5.3 million users and 180 million unique bookmarked resources. The exact number of users when Delicious was re-sold to AVOS on 2011 has not been found in the literature.
Our data set was crawled from November, 2009 to January, 2010, then again from May, 2010 to August, 2010. The data set is described in Table I

Table 1: Data Statistics

Total resources	41,469,488	**Total tags**	214,420,801
Total bookmarks	73,216,330	**Total users**	723,342
Min bookmarks	1	**Max bookmarks**	9099
Avg bookmarks per resource	2	**Avg tag per bookmark**	3

4 Data Analysis

Using the dataset described above, various issues addressed or discussed in the literature, either explicitly or implicitly were examined. The goal was to verify what had been reported in the literature or qualify those findings. The issues explored included the influence of early tagging behavior, the propagation of bookmarks on resources, and the adoption of dominant tags. Each of these issues is addressed below.

4.1 At what rate do bookmarks accumulate?

Golder and Huberman [7] noted that among their 212 most popular resources they found that most resources receive their bookmarks in the first 10 days while some others stay in the system for at least six month before they experience a burst in popularity. They observed that all resources experience "peak in popularity". We investigated those pattern and found that both patterns exists on our data set "Early and Late adoption" but we have also found one more "Gradual pattern".

4.1.1 Gradual Pattern

In this pattern, the relationship between the number of bookmarks associated with a resource and time is generally linear. The resource accumulates bookmarks gradually with time. This indicates ongoing attention from users. Figure1 shows this linear relationship. Other than the users' interest, this also can be an indication of the type of this resource. In Figure 1 the resource is the search engine 'Bing', which is a resource that might be used by users on a regular basis and whose popularity is increased through recommendations of users

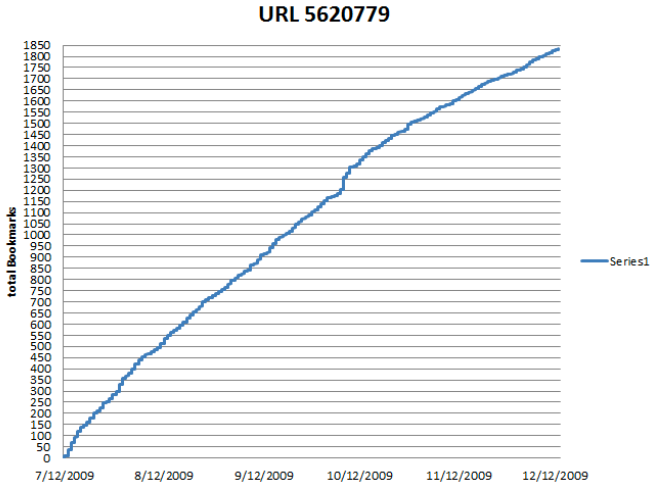


Figure 1: Gradual Bookmarks accumulation

4.1.2 Early adoption

Under this pattern a resource accumulates bookmarks very quickly, in some cases a thousand bookmarks in one day, and then a very slow growth rate ensues (see figure 2). These kinds of resources attract attention instantly. We suspect that external events may be more important here. For example this can be a news article or an important event. This pattern was confirmed by [7, 8]

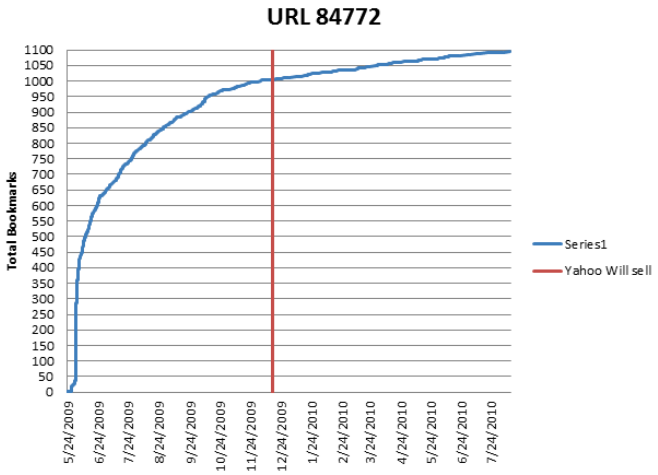


Figure 2: Early Adoption

4.1.3 Late adoption

Under this pattern a resource is introduced and not bookmarked frequently. Suddenly, attention is attracted to it (Figure 3). This pattern can be an indication of many things: a sudden increase in the importance of that resource or external events that might lead to the popularity of this resource.

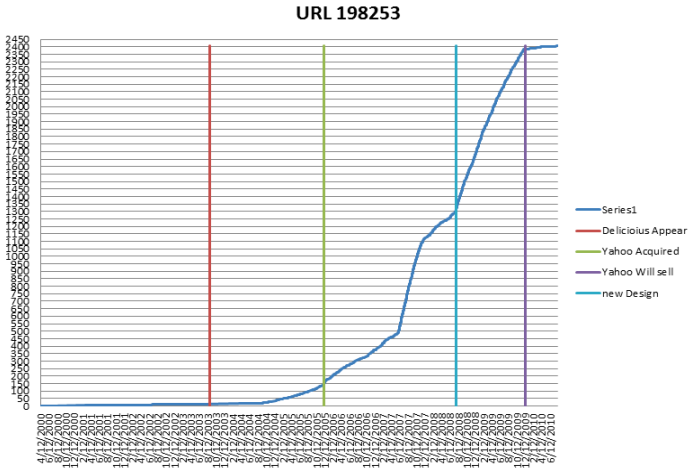


Figure 3: Late Adoption

Note that Figure 2 and Figure 3 shows a set of events related to Delicious as a system. There is some indication that some resources grow in popularity related to these system events.

4.2 To what extent do early tags influence later tagging behavior?

Golder and Huberman [7] found that users tend to use more general tags first and those general tags are the most frequently used meaning that the tagging behavior of later users is heavily influenced by early taggers. They also pointed out that after 100 bookmarks the pattern of tagging stabilizes due to imitation and shared knowledge[7]. While this current study is generally supportive of the Golder and Huberman findings, the actual structure is somewhat more complex. To examine how tagging evolved over time, we looked at the resource at different points in its lifetime (a) one week (b) one month (c) three months (d) six months (e) one year and (f) the most recent date for which we have information (as little as one year and as much as six years). We also considered the tag set at this point as a mature set. The accumulation of tags over time was visualized as shown in Figure 4 where the resource is represented by a blue circle in the middle and each tag is placed on a spiral ordered by creation date moving clockwise. The size of the triangle increases as the frequency of use increases. Tags that have been used more than once are labeled. Running this visualization on a sample, we found a number of different patterns illustrated in the next section

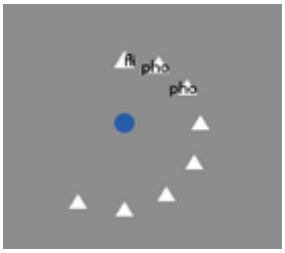


Figure 4: tagging pattern visualization

4.2.1 Early Tag Dominance Pattern:

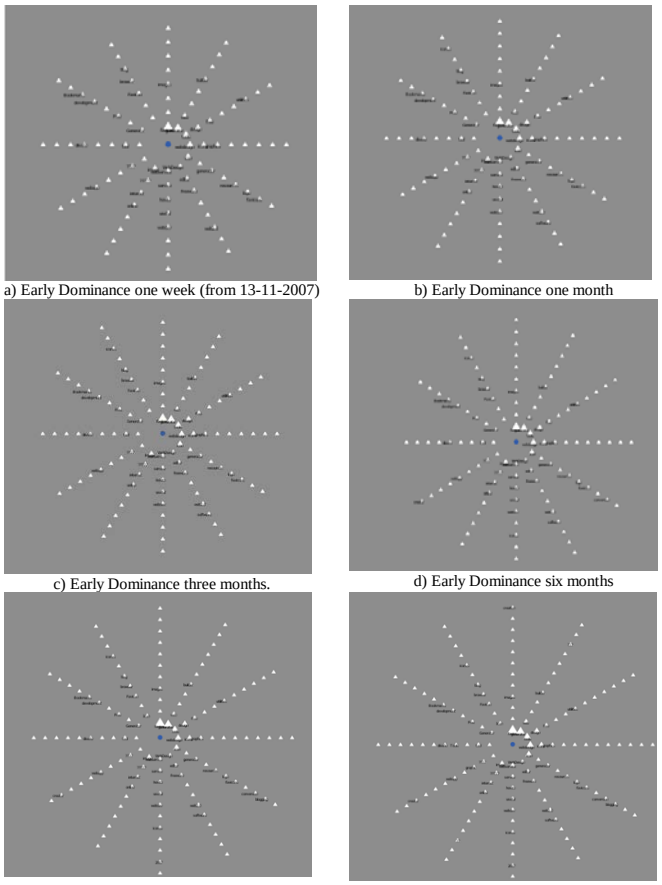
In this pattern, as the bookmarks accumulate for that resource, early tags are used most frequently. If we consider the popularity of a tag by its usage, earlier tags are more popular in this pattern. This pattern was first described in [7] and it is illustrated in (Figure 5a-5f). The selected resource accumulates bookmarks slowly and the increase in the tags usage is gradual so that we can see at the end that early tags are the most popular. The analysis shows this pattern to be far from the dominant pattern. Over a sample of 53 resources, this pattern was observed about 49% of the time.

4.2.2 Mixed Tag Dominance Patterns

In this pattern, as the resource accumulates more bookmarks, we see that both early and later tags are dominant. This pattern is illustrated in Figure 6. In other cases it is also possible that the later tags are the most dominant.

This pattern is illustrated in Figure 6. The first bookmark that belongs to this resource was added to delicious on 02-11-2004. After one week this resource accumulated 8 tags with usage frequency of 1 for each tag. A month later, there were no changes in this pattern. Three months later (Figure 6b), two new tags were added. After six months (Figure 6c), the tag usage started to increase. Then, at one year, more tags accumulated and the tag usage has increased. At the last known bookmark for this resource on our data set (10-8-2010), the most popular tag is 'programming' which was

added to the system at about three months as were other popular tags (Figure 6e). Variations of this pattern were observed for 38% of the 53 resources that were examined.



a) Early Dominance one week (from 13-11-2007) b) Early Dominance one month

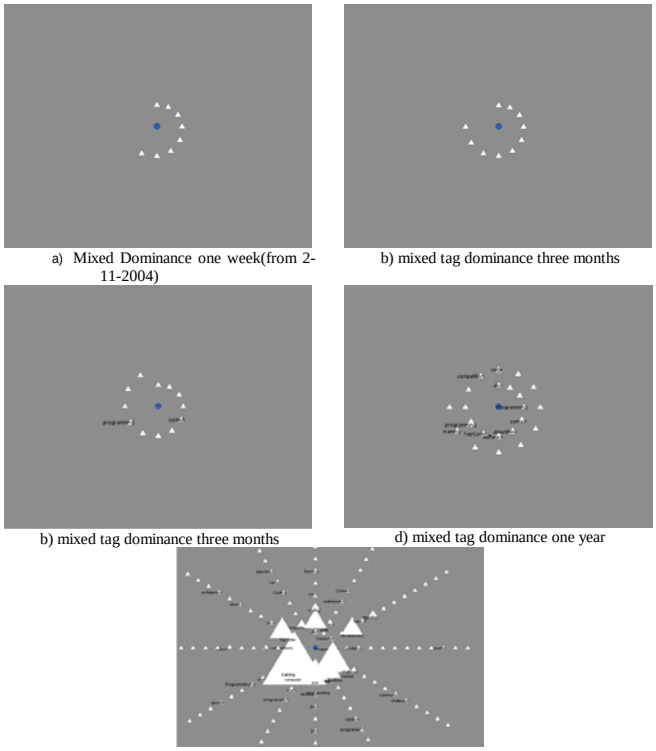
c) Early Dominance three months. d) Early Dominance six months

E) Early Dominance one year. F) Early dominance last Known bookmark(11-8 - 2010)

Figure 5: Early tag dominance pattern

4.2.3 Rapid versus slow growth patterns

As illustrated in Figure 7, some resources grow rapidly accumulating a large set of tags from their first day in the system while other start with few bookmarks. Figure 7 shows two snap shots of two different resources after one week from the day they were first added to the system. The resource on the left has about 100 bookmarks at the end of the first week while the resource on the right has two bookmarks with 8 tags. The visualization on their tag set at the latest date shows that both resources develop a similar pattern even though one grew slowly and the other rapidly.

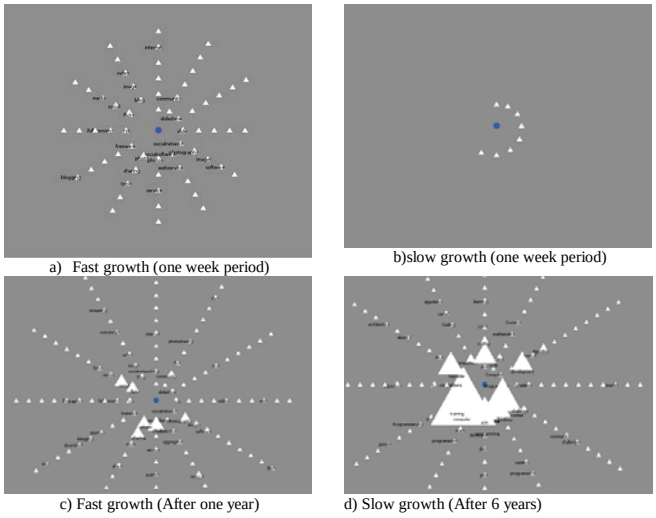


a) Mixed Dominance one week(from 2-11-2004) b) mixed tag dominance three months

b) mixed tag dominance three months d) mixed tag dominance one year

e) mixed tag dominance - last known bookmark (10-8-2010)

Figure 6: Mixed tag dominance pattern



a) Fast growth (one week period) b)slow growth (one week period)

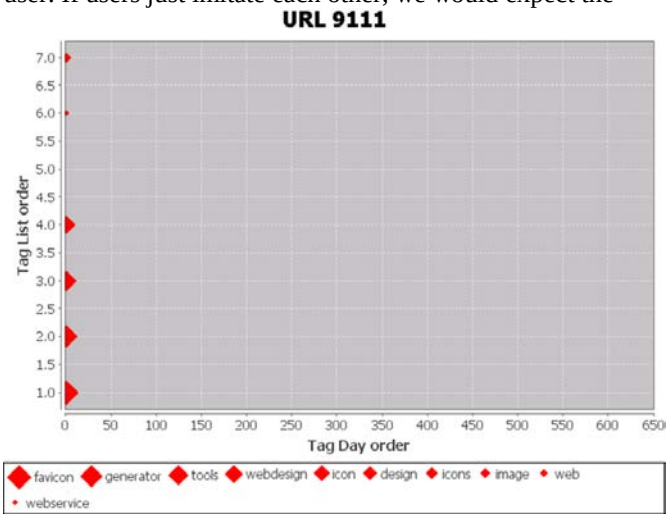
c) Fast growth (After one year) d) Slow growth (After 6 years)

Figure 7: Fast vs slow growth

Golder and Huberman [7] hypothesized that the order of tag growth is influenced by imitation and shared knowledge. Other studies followed that tried to build a model to simulate the behavior of those systems focusing on imitation as the only factor that affected the behavior [9, 13]. Another study [14] built a model in which they simulated the effect of imitation and shared knowledge, their model performed better than the previous models. We show here that the

underlying structure may have more factors that influence the behavior.

Figure 8 shows that the growth of the top 10 tags were influenced by imitation. The first four tags that were added to the system were ranked by their order (1st, 2nd, 3rd and 4th) and all the tags that were added to the system on the first day were the most popular tags. Note that 'webdesign', 'icon' and 'design' were all added to the system on the first day and they were fourth on the sequence and their crosses on the chart are overlapping. Figure 9 shows the timing order for top10 tags for the resource www.freelance.com. We can see that the most frequently used tag was added to the system on day 133 and it was the second on the list of tags provided by the user 'finrod'. We can also see that 'outsourcing' was added to the system on day 266 and it was ranked 3rd. One would expect that the user would provide more descriptive tags first but surprisingly, 'outsourcing' was the seventh on the list of tags provided by the user and it was ranked 3rd. Furthermore, Delicious provides a list of seven recommended tags which are the most popular for that resource and the seven most frequently used tags for that user. If users just imitate each other, we would expect the

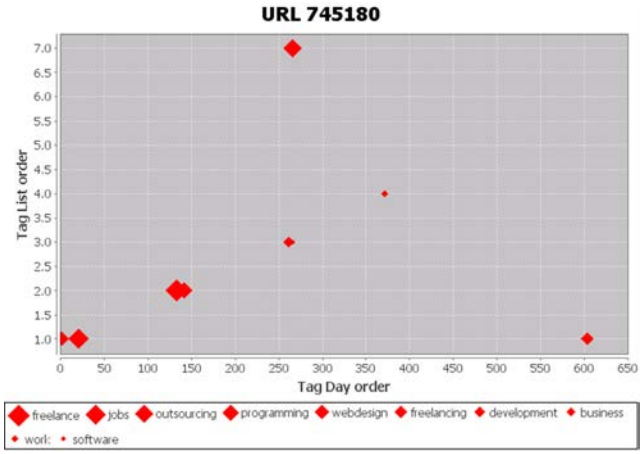


Rank	Tag	Day order	List Order	Frequency	User
1	**Favicon**	1	1	152	bargainbinburglar
2	**Generator**	1	2	107	bargainbinburglar
3	**Tools**	1	3	82	bargainbinburglar
4	**webdesign**	1	4	82	bargainbinburglar
5	**Icon**	1	4	58	biscotheque
6	**Design**	1	4	53	conque
7	**Icons**	1	2	47	ombran
8	**Image**	1	7	34	biscotheque
9	**Web**	1	1	31	Conque
10	**Webservice**	1	6	23	biscotheque

Figure 8: Top 10 tags ordered by time

Tagging behavior to be somewhat similar to Figure8 but Figure 9 shows a totally different behavior.

5) Conclusions about Tagging Patterns: Resources start with a set of tags. The most popular tags may be from the initial offering or they may appear later. When a resource attracts a lot of early attention there is even less chance that dominant tags will be the first used. Clearly, there are many factors influencing the tagging pattern other than the influence of early tags. Does it mean that each group of users perceives the resource differently or is it a change in the users' interests? To further answer this question, we are planning an examination of differences between the semantics of tags that become dominant and those that do not.



Rank	Tag	Day Order	List Order	Frequency	User ID
1	freelance	133	2	959	Finrod
2	jobs	20	1	526	Scazza
3	outsourcing	266	7	383	edelwater
4	programming	141	2	333	Afriza
5	webdesign	1	1	270	finnstones
6	freelancing	604	1	245	butamt
7	development	261	3	224	chulan
8	business	2	1	223	b3n1tora
9	work	371	4	167	altoven
10	software	266	3	134	edelwater

Figure 9 Top 10 tags ordered by time

4.3 How do the top 10 tags evolve?

We define the top 10 tags as the 10 most frequently used tags regardless of when they were introduced in the system. We inspected the frequency of each of those tags at various points in time – every 30 days. In general we observed two

patterns: exponential popularity growth and linear popularity growth. The exponential pattern where a tag stays unpopular for a while and then its popularity explodes (Figure 10). On the other hand, the linear pattern experiences a sudden growth at the beginning and then a stable growth (Figure 11). The exponential pattern corresponds to the early adoption pattern found in the bookmark accumulation while the linear pattern corresponds to the late adoption. We tried to relate those tag growth patterns to the observed bookmarking patterns (Section IV-A). We found that gradual and late adoption patterns tend to follow the exponential growth pattern while the early adoption pattern tends to follow the linear growth pattern. This is not to generalize the case because it is not true all the time but it is to raise the point that social bookmarking system behavior needs further investigation. Sometimes the top 10 tags will start to grow exponentially and then stabilize to a linear pattern of growth. We are clearly in need for a further understanding of these different cases so we can make more coherent conclusions on those tags.

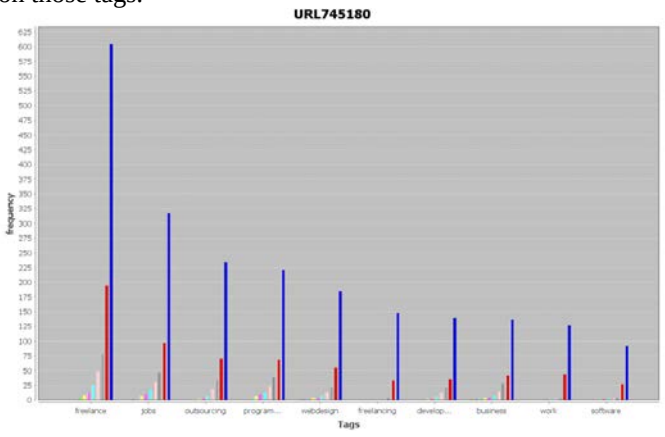


Figure 10: Exponential Popularity

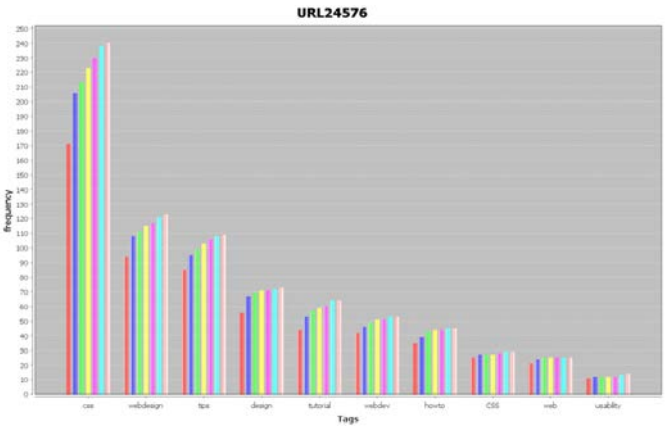


Figure 11:Linear Popularity

5 Discussion:

This paper reexamined some of the assumptions about how users interact with resources in social bookmarking systems. This analysis examined how resources accumulate tags. We found that some resources accumulate tags rapidly while others accumulate them slowly? .Which tags become the most popular was also examined. In folksonomies, tags are measures of agreement among users. It is not necessarily the case that early tags are the most dominant. If the early tags are the most dominant, it may simply be a matter of users imitating early tags or it may be the case that there is agreement among users. When tags that appeared later become the most dominant it may indicate a disagreement among users, or it may be the case that users are more discriminating about which are the correct tags. This implies that the tagging practice is not as influenced by early taggers or Delicious' suggested tags as some have suggested [7, 8].

Analyzing how bookmarks accumulate overtime, we found that different behaviors may attribute different meanings. A resource that attracts attention gradually might not be as useful as a resource which attracted attention early on. For example: in searching for extreme events such as natural disasters we might need to consider a resource that follows the early adoption pattern. On the other hand, searching for a popular restaurant or popular shopping sites we might want to consider those resources that grow gradually because those sites popularity might grow slowly through recommendations by users.

Both bookmark and tag evolution patterns need to be considered when using SBSs. For example, if we are using tags for ranking, we should consider how tags evolve and how the resource accumulates bookmarks. Furthermore, when ranking documents, if we consider only early popular tags, we might be losing an important description that can be obtained from the later popular tags and vice versa. Similarly, tag accumulation rate should also be considered. A resource that accumulates tags early should not be treated the same as a resource which accumulated tags slowly.

We were able to confirm some of the findings in [7] and [8] in what we call an early adoption pattern. Wetzker et al [8] states that "Delicious community pays attention to new resources only for a very short period of time. As a result, these resources receive most of their posts very quickly and disappear shortly afterwards". We confirmed the existence of this pattern as well as other patterns such as the late adoption patterns described in [7] in which a resource experiences a sharp growth after a period of suppression. We have also found that other resources experience gradual growth over time which might be due to continued user interest.

We found that there are different possible scenarios of how those tags grow over time. There are two clear patterns: exponential and linear but we can't generalize the finding without further investigation of those patterns and their meanings. A tag that grows linearly might mean a rapid agreement between users while a tag that grows exponentially means a sudden change in this agreement. The growth of late tags might as well be a combined effect of imitation and shared knowledge in which a user might add all the recommended tags first and then add their tags which can attract more users later.

6 Conclusions

This paper examined how tags accumulate over time and how those tags grow in popularity. We have confirmed the existence of the patterns described in early research [7-9, 12] but we have also observed other patterns that were not discussed in the literature which might be important. It would be useful to examine the semantics of dominant tags, potentially through a time analysis of when specific semantics become dominant. Also, more careful analysis of the cognitive aspects of user behavior might be useful in understanding the semantics of the list of tags provided by the user to describe a specific resource.

Research reported in this publication was supported by Saudi Arabian Cultural Mission to the U.S.

7 References

[1] T. Hammond, T. Hannay, B. Lund, and J. Scott, "Social bookmarking tools (I) a general review," *D-lib Magazine*, vol. 2, 2005.

[2] T. V. Wal. (2007, 5/21). *Folksonomy*. Available: http://vanderwal.net/folksonomy.html

[3] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Can social bookmarking enhance search in the web?," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 107-116.

[4] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 193-202.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 195-206.

[6] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 501-510.

[7] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of information science*, vol. 32, pp. 198-208, 2006.

[8] R. Wetzker, C. Zimmermann, and C. Bauckhage, "Analyzing social bookmarking systems: A del. icio. us cookbook," in *Proceedings of the ECAI 2008 Mining Social Data Workshop*, 2008, pp. 26-30.

[9] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 211-220.

[10] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proceedings of the seventeenth conference on Hypertext and hypermedia*, 2006, pp. 31-40.

[11] U. Farooq, T. G. Kannampallil, Y. Song, C. H. Ganoe, J. M. Carroll, and L. Giles, "Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics," in *Proceedings of the 2007 international ACM conference on Supporting group work*, 2007, pp. 351-360.

[12] M. E. Kipp and D. G. Campbell, "Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices," *Proceedings of the American Society for Information Science and Technology*, vol. 43, pp. 1-18, 2006.

[13] C. Cattuto, V. Loreto, and L. Pietronero, "Collaborative tagging and semiotic dynamics," *arXiv preprint cs/0605015*, 2006.

[14] K. Dellschaft and S. Staab, "An epistemic dynamic model for tagging systems," in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, 2008, pp. 71-80.

[1]

The Use of Merging Algorithm to Real Ranking for Graph Search

A. Mohammad Reza Nami , B. Mehdi Ebadian

Faculty of Electrical, Computer, and IT Engineering,
Islamic Azad University- Qazvin Branch, Qazvin, IRAN

ABSTRACT

Ranking problem is becoming an important issue in many fields, especially in information retrieval. This paper presents an automatic technique for spam monitoring in the graph. The technique is based on combining information from two different sources: Truncated page rank and Semi-Streaming Graph Algorithms. In this paper we conduct further study on the heuristically ranking framework and provide measuring page rank of link farm. Twenty-six articles from 15 venues have been reviewed and classified within the taxonomy in order to organize and structure existing work in the field of Information Retrieval.

Keywords

Information retrieval (IR), Page rank (PR), Streaming Algorithms, Internet Marketing, Spam and Search Engine Optimization.

1 Introduction

Search engines have being become the most lucrative thing over the internet. Search engines are mediated between Web platform and information seeker. Search engines then rank Web pages to create short list of high-quality result. On the other hand, large visits originate from search engines that most users just click on first few results. Therefore, creating high score page independently of their real merit.

SPAM: Each new communication Media creates opportunity for sending unsolicited messages. Type of electronic spam includes e-mail spam, instant messaging (SPIM), internet telephony (SPIT), spamming by mobile phone, by fax, and so on. The request responses paradigms of HTTP so goal is deceive search engines.

Any attempt to deceive search engine's relevancy algorithm or "would not be done if search engines did not exist" So ethical attempt is different between SPAM and SEO (Search Engine Optimization) . The relation between website and search engine administrator is adversarial.

Stream graph algorithm: Suppose that we have a very large undirected, un-weighted graph (starting at hundreds of millions of vertices, ~10 edges per vertex), non-distributed and processed by single thread only and that I want to do breadth-first searches on it. I expect them to be I/O-bound, thus I need a good-for-BFS disk page layout, and disk space is not an issue. The searches can start on every vertex with equal probability. Intuitively that means minimizing the number of edges between vertices on different disk pages, which is a graph partitioning problem. The graph itself looks like spaghetti think of random set of points randomly interconnected, with some bias towards shorter edges.

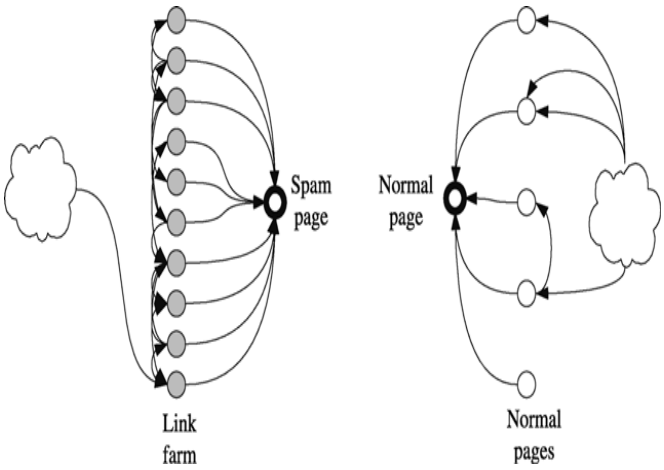


Figure 1. Link farm (Link-Base Web Spam (Topological Spam))

Web spam techniques classified two groups: content (keyword) spam, and link spam.

Link spam changes the sites structure by creating link farm.

Link farm is densely connected pages to deceiving ranking algorithm by improving one user in group.

Our spam-detection algorithm target are pages which receive most link-base ranking by participating in link farms but little relationship with rest of the graph.

Links may not be spam, by buying advertising or buying expired domains that used legitimate purposes.

Topological spamming is spamming which achieved by using Link farm.

Link-based and content-based analysis offers two orthogonal approaches. Weakness of link-based: For some pages that statistically close to non spam pages.

Threats of link -based: Hybrid spam structure.

Opportunity of link-based: Link farms are expensive.

Weakness of content -based: less resilient to changes in spammer strategies.

Threats of content -based: Hybrid spam structure, copy entire Web site (change few out-link) is inexpensive.

So they should be used together.

2 Algorithm Framework

Fetterly et al. [2004] hypothesized statistical distribution about pages is a good way to detecting spam pages, "in a number of these distribution, outlier values are web spam".

Baeza-Yates et al. [2006] introduce damping function for rank propagation.

We want to explore the neighborhood of page and link structure artificially generated or not.

Two algorithm challenges:

1. how to simultaneously compute statistics
   neighborhood of each page in huge web graph
2. how use it to detect and demote web spam

2.1 Supporter

If there is a link page x to page y, the author of page x is recommending page y, the x is supporter of page y at distance d, if shortest path from x to y formed by links in E has length d .

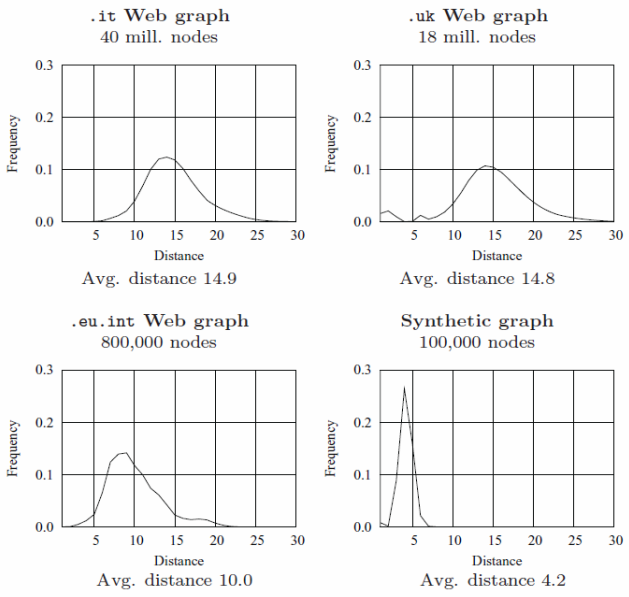


Figure 2. Web Graph and supporter Distribution.

Distribution of the fraction of distinct supporters found at varying distances (normalized), obtained by backward breadth-first visits from a sample of nodes, in four large Web graphs.

Number of new distinct supporter increases up to certain distance, and the decreases, graph is limit in size and we approach effective diameter.

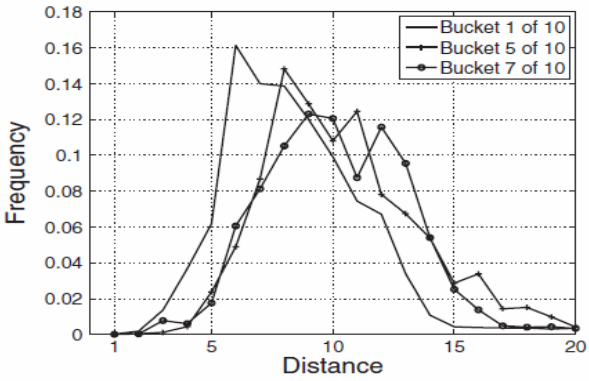


Figure 3. Different Bucket's page ranks.

Calculate Page Rank (PR) of pages in the eu.int sub domain to showing different distribution in high and low ranked sites.

Breadth-first search (BFS) instead of computing the distribution for all nodes of sample of large Web graphs.

Advantage: inexpensive

Disadvantage: memory for each marked nodes $\Omega(N^2)$ time to repeat BFS.Solution: compute supporters only for subset of suspicious nodes
constraint: we do not know a prior node is suspicious of being spam or not.

```
Require: graph G = (V, E), score vector S
INITIALIZE(S)
while not CONVERGED do
  for src : 1 ... |V| do
    for all links from src to dest do
      COMPUTE(S, src, dest)
    end for
  end for
  POST_PROCESS(S)
end while
return S
```

Algorithm 1: Link-analysis algorithm

Link-analysis algorithm using semi-stream model, metric is score vector that uses O(N log N) bits memory.

PR algorithm instead of BFS for web spam detection, for measure the centrality of nodes outcomes tree a specific node and not all nodes, whereas PR compute a score for all nodes in the graph at same time.

2.2 TRUNKATED PAGERANK

A link-based ranking function that reduces importance of neighbors which topologically close to the target node.

Damping function ignores direct contribution of the first levels of links. Spam pages should be very sensitive to changes in damping factor of PR calculation.

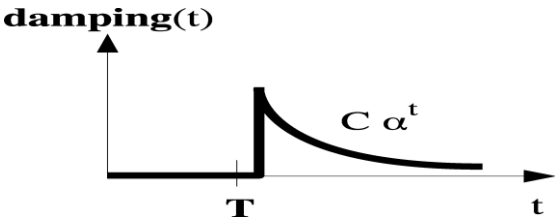
$A_{N\times N}$ be citation matrix of

$$G = (V, E), \quad \alpha_{xy} = 1 \Leftrightarrow (x, y) \in E \qquad (1)$$

P be row-normalized citation matrix, that all rows sum up to one, and rows of zeros replaced 1/N to avoid sink rank.

$$W = \Sigma [damping(t) / N] P^t$$

$$\text{Damping}(t) = \begin{cases} 0 & t \le T \\ C\alpha^t & t > T \end{cases} \qquad (2)$$



C is normalization constant

α is damping factor

```
Require: N: number of nodes, 0 < α < 1: damping factor, T ≥ −1: truncation
  distance
for i : 1 ... N do {Initialization}
  R[i] ← (1 − α)/((α^(T+1))N)
  if T ≥ 0 then
    Score[i] ← 0
  else {Calculate normal PageRank}
    Score[i] ← R[i]
  end if
end for
distance = 1
while not converged do
  Aux ← 0
  for src : 1 ... N do {Follow links in the graph}
    for all link from src to dest do
      Aux[dest] ← Aux[dest] + R[src]/outdegree(src)
    end for
  end for
  for i : 1 ... N do {Apply damping factor α}
    R[i] ← Aux[i] ×α
    if distance > T then {Add to ranking value}
      Score[i] ← Score[i] + R[i]
    end if
  end for
  distance = distance +1
end while
return Score
```

Algorithm 2

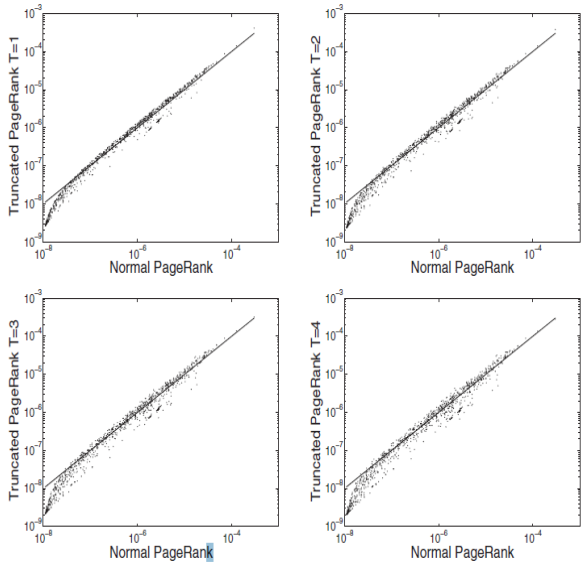


Figure 4. 4times truncated page rank.

With comparing PR and TPR, for value from 1 to 4, both closely correlated, an correlation decreases as more level truncated.

2.3 ESTIMATION SUPPORTERS

Use probabilistic counting to compute estimation the number of supporter for all vertices in the graph at the same time.

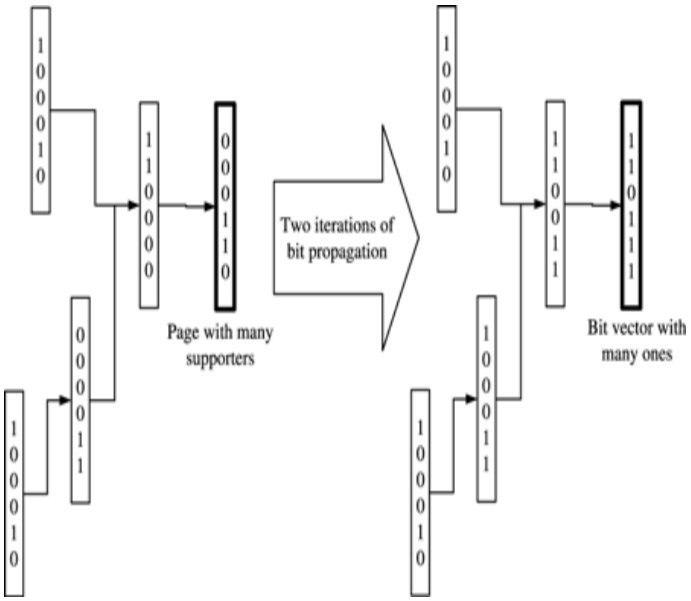


Figure 5. Propagation of having supporter 1 and Not 0 .

Bit propagation algorithm. Page y has a link to page x, then vector of page x is updated: x ← x OR y

Bit propagation Algorithm for estimating number of distinct supporters at distance ≤ d of all nodes.

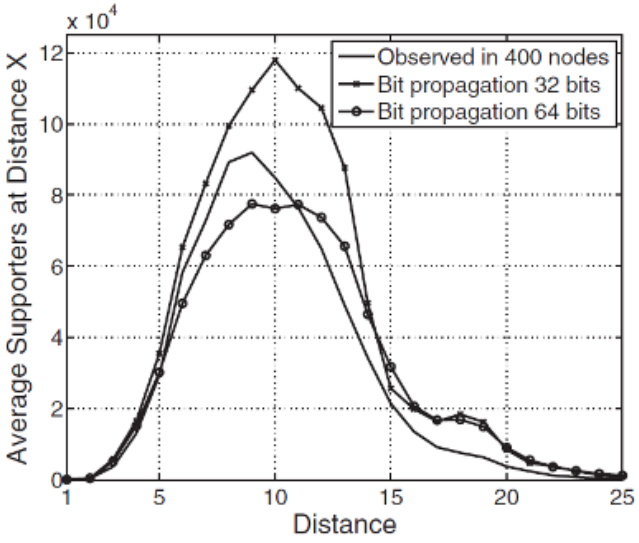


Figure 6. Distances of supporter in 3 types.

Comparison of estimation average number of supporters against observed value in a sample of nodes, by assuming

$\epsilon = 1/N$

$$\overline{N}(x) = \log_{(1-\epsilon)}\left(1 - \frac{\overline{B}_\epsilon(x)}{k}\right) \quad (3)$$

Require: N: number of nodes, d: distance, k: bits
1: **for** node : 1 ... N **do** {Every node}
2: **for** bit : 1 ... k **do** {Every bit}
3: INIT(node,bit)
4: **end for**
5: **end for**
6: **for** distance : 1 ... d **do** {Iteration step}
7: Aux ← $\mathbf{0}_k$
8: **for** src : 1 ... N **do** {Follow links in the graph}
9: **for all** links from src to dest **do**
10: Aux[dest] ← Aux[dest] OR V[src,·]
11: **end for**
12: **end for**
13: **for** node : 1 ... N **do**
14: V[node,·] ← Aux[node]
15: **end for**
16: **end for**
17: **for** node: 1 ... N **do** {Estimate supporters}
18: Supporters[node] ← ESTIMATE(V[node,·])
19: **end for**
20: **return** Supporters

Table 1. Performance of this Article classifier

Metrics	UK2012			UK2013		
	True Positive	False Positive	F1	True Positive	False Positive	F1
Degree (D)	0.733	0.016	0.807	0.324	0.023	0.431
D + Page Rnk (P)	0.769	0.014	0.836	0.36	0.026	0.467
D+P +Trust Rank	0.785	0.013	0.847	0.54	0.038	0.596
D + P+ Trunc. PR	0.782	0.016	0.844	0.356	0.021	0.474
D + P +Est. Supporters	0.801	0.008	0.868	0.467	0.038	0.549
All attributes	0.806	0.01	0.872	0.586	0.038	0.632

And Estimation with adaptive Bit propagation, by dividing ϵ two at each iteration b

3 Classification

- Precision P = tp/(tp + fp) → P = #spam hosts classified as spam /(#hosts classified)
- Recall R = tp/(tp + fn) → R = #spam hosts classified as spam/(#spam hosts)
- Fp → False positive rate = #normal hosts classified as spam / (#normal hosts)
- Fn → False negative rate = # spam host classified as spam / (#spam hosts)

Table 2. Criterion "F" (Web spam techniques classification)

	Relevant Spam hosts	Nonrelevant Normal hosts
Retrieved	tp #spam hosts classified as spam	fp
Not Retrieved	fn	tn #normal hosts not classified as spam

Table 3. Performance Using Page Rank Supporters degree Experimental Result

Dataset	True Positive	False Positive	F-Measure	Previouse F-Measure from Table IV
UK	0.801	0.008	0.866	0.834
Only pages	0.795	0.014	0.853	
Only hosts	0.778	0.011	0.849	
UK	0.465	0.033	0.549	0.459
Only pages	0.402	0.03	0.497	
Only hosts	0.468	0.03	0.555	

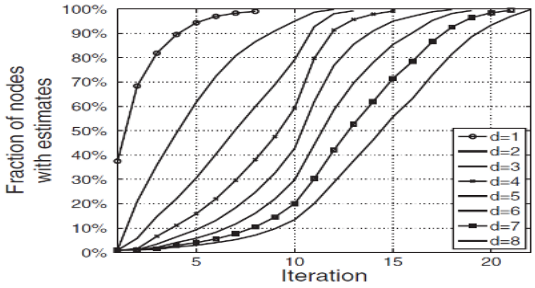


Figure 7. Best Iteration to find suitable distance

4 Conclusions

The technique used for link analysis assigns to every node in Page Rank the web graph a numerical score between 0 and 1, known as its Page Rank. With the help of this paper the website owners and webmasters can decide which SEO practice is worth and will give a good return on investment.

Finally, the use of regularization methods that exploit the topology of the graph and the locality hypothesis [Davison 2000b] is promising, as it has been shown that those methods are useful for general Web classification tasks [Zhang et al. 2006; Angelova and Weikum 2006; Qi and Davison 2006] and that can be used to improve the accuracy of Web spam detection systems [Castillo et al. 2007].

REFERENCES

[1] Alexa Inc., http://www.alexa.com/help/traffic-learn-more last accessed on may 17, 2011

[2] Antoniol, G. and Guéhéneuc, Y. G., "Feature Identification: An Epidemiological Metaphor", IEEE Transactions on Software Engineering, vol. 32, no. 9, 2006, pp. 627-641.

[3] Binkley D, Gold G, Harman M, Li Z, Mahdavi K (2008) An empirical study of the relationship between the concepts expressed in source code and dependence. J Syst Software 81:2287–2298

[4] Cornelissen B, Zaidman A, van Deursen A, Moonen L, Koschke R (2009) A systematic survey of program comprehension through dynamic analysis. IEEE Trans Software Eng (TSE) 35(5):684–702

[5] De Alwis B, Murphy GC (2008) Answering conceptual queries with Ferret. 30th International Conference on Software Engineering (ICSE'08), Leipzig, Germany, 21–30

[6] De Lucia, A., Fasano, F., Oliveto, R., and Tortora, G., "Recovering Traceability Links in Software Artefact Management Systems", ACM Transactions on Software Engineering and Methodology, 2007.

[7] Egyed, A., Binder, G., and Grunbacher, P., "STRADA: A Tool for Scenario-Based Feature-to-Code Trace Detection and Analysis", in Proc. of IEEE/ACM 29th International Conference on Software Engineering (ICSE'07), 2007, pp. 41-42.

[8] Eaddy M, Aho AV, Antoniol G, Guéhéneuc YG (2008a) CERBERUS: tracing requirements to source code using information retrieval, dynamic analysis, and program analysis. 16th IEEE International Conference on Program Comprehension (ICPC'08), Amsterdam, The Netherlands, 53–62

[9] Eaddy M, Zimmermann T, Sherwood K, Garg V, Murphy G, Nagappan N, Aho AV (2008b) Do crosscutting concerns cause defects? IEEE Trans Software Eng 34(4):497–515

[10] Gay G, Haiduc S, Marcus M, Menzies T (2009) On the use of relevance feedback in IR-based concept location. 25th IEEE International Conference on Software Maintenance (ICSM'09), Edmonton, Canada, 351–360

[11] Grant S, Cordy JR, Skillicorn DB (2008) Automated concept location using independent component analysis 15th Working Conference on Reverse Engineering (WCRE'08), Antwerp, Belgium, 138–142

[12] Hayes, J. H., Dekhtyar, A., and Sundaram, S. K., "Advancing candidate link generation for requirements tracing: the study of methods", IEEE Transactions on Software Engineering, vol. 32, no. 1, January 2006 2006, pp. 4-19.

[13] Hill E, Pollock L, Vijay-Shanker K (2009) Automatically capturing source code context of NL-queries for software maintenance and reuse. 31st IEEE/ACM International Conference on Software Engineering (ICSE'09), Vancouver, British Columbia, Canada

[14] Kothari, J., Denton, T., Mancoridis, S., and Shokoufandeh, A., "On Computing the Canonical Features of Software Systems", in 13th IEEE Working Conference on Reverse Engineering (WCRE'06), Benevento, Italy, 2006.

[15] Kuhn, A., Ducasse, S., and Gîrba, T., "Semantic Clustering: Identifying Topics in Source Code", Information and Software Technology, vol. 49, no. 3, March 2006, pp. 230-243.

[16] Lawrance J, Bellamy R, Burnett M (2007) Scents in programs: does information foraging theory apply to program maintenance? IEEE Symposium on Visual Languages and Human-Centric Computing (VL/ HCC'07), IEEE, 15–22

[17] Liu D, Marcus A, Poshyvanyk D, Rajlich V (2007) Feature location via information retrieval based filtering of a single scenario execution trace. 22nd IEEE/ACM International Conference on Automated Software Engineering (ASE'07), Atlanta, Georgia, 234–243

[18] Li Z (2009) Identifying high-level dependence structures using slice-based dependence analysis. King's College London, University of London. Ph.D

[19] Lukins S, Kraft N, Etzkorn L (2008) Source code retrieval for bug location using latent dirichlet allocation. 15th Working Conference on Reverse Engineering (WCRE'08), Antwerp, Belgium, 155–164

[20] Poshyvanyk, D., Guéhéneuc, G. Y., Marcus, A., Antoniol, G., and Rajlich, V., "Feature Location using Probabilistic Ranking of Methods based on Execution Scenarios and Information Retrieval", IEEE Transactions on Software Engineering, vol. 33, no. 6, June 2007, pp. 420-432.

[21] Rajlich, V., "Changing the Paradigm of Software Engineering", in Communications of ACM, vol. August, 2006, pp. 67-70.

[22] Salah, M., Mancoridis, S., Antoniol, G., and Di Penta, M., "Scenario-driven dynamic analysis for comprehending large software systems", in Proc. of 10th European Conference on Software Maintenance and Reengineering (CSMR'06), 2006.

[23]Shepherd, D., Fry, Z., Gibson, E., Pollock, L., and Vijay-Shanker, K., "Using Natural Language Program Analysis to Locate and Understand Action-Oriented Concerns", in Proc. of International Conference on Aspect Oriented Software Development (AOSD'07), 2007, pp. 212-224.

[24] Simmons, S., Edwards, D., Wilde, N., Homan, J., and Groble, M., "Industrial tools for the feature location problem: an exploratory study", Journal of Software Maintenance: Research and Practice, vol. 18, no. 6, 2006, pp. 457-474.

[25]WordStreamTools, http://www.wordstream.com/adwordskeyword-tool on May 10, 2011

[26] Zhao, W., Zhang, L., Liu, Y., Sun, J., and Yang, F., "SNIAFL: Towards a Static Non-interactive Approach to Feature Location", ACM Transactions on Software

Engineering and Methodologies, vol. 15, no. 2, 2006, pp. 195-226.

Keyword Searches with Customized Preferences

Yu-Chin Liu, Yi-Hsuan Chiang and Yu-Lien Hsieh

***Abstract*—In accordance with the great business opportunities emerging through SNSs, entrepreneurs strive to explore the potential benefits by analyzing data collected from SNSs. For example, Google+ attempts to integrate keyword searches with the individual's social network. In this paper, we propose a new method for considering the common preferences of friends on social networks while ranking the order of related web pages returned from search engines. The simulation shows the proposed method performing well comparing to general search engines.**

I. Introduction

INFORMATION searching has become as one of the most important tasks for on-line information retrieval. At present, there are three main methods of information searching on the Internet: searching by web pages, by directories, and by keywords. For users, tools supporting timely searching are in great demands.

As stated by [1], people commonly first input keywords on search engines; and the search engines try hard to find the web pages and rank the related pages in descending order of significance. Finally, people browse the ranked web pages to find the answers they desire.

The technique of keyword searches on the Internet is like an extension of information retrieval from documents; with keyword searching, people can garner required information from on-line web pages at relatively low costs [2].

[3] and SEO (Search Engine Optimization) practitioners have found that most information seekers will not click URLs listed beyond top three pages. Commonly, 75% of the URLs listed on the first page will be browsed. Once people find these links are not working properly, they will change the keywords to initiate another search or try other search engines. Therefore, the techniques for ranking related URLs correctly are essential to search engines.

Existing search engines usually compute the similarities between keywords provided by users and related web pages to order the returned URLs by their descending similarities to the requests. In real cases, the ranks of returned URLs do affect the effectiveness and efficiency of information searching. Hence, our paper proposes a new method for customizing URL page ranks for query results by incorporating the common preferences within social groups. To the best of our knowledge, as the development of Social Network Sites is relatively young, there is no research ranking returned URLs of query results in terms of personalized considerations. Our research uses the data provided by InsightXplorer which logs users' browsing paths.

The rest of the paper is organized as follows: the related work is reviewed in Section 2, while Section 3 details the proposed methods. Section 4 considers several experiments to verify the effectiveness of the proposed method by comparing it with Google Search results, and the conclusions are made in Section 5.

This work was supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the contract number of NSC 102-2410-H-128-030-MY2.

Y. C. Liu is with the Department of Information Management, Shih Hsin University, Taipei, Taiwan 11604, ROC. (phone: 886-2-22368225-3362; fax: 886-2-2236-7114; e-mail: ycliu@mail.shu.edu.tw).

Y. H. Chiang is with the Department of Department of Radio, TV, & Film, Shin Hsin University, Shih Hsin University, Taipei, Taiwan 11604, ROC.

Y. L. Hsieh is with the Department of Information Management, Shih Hsin University, Taipei, Taiwan 11604, ROC.

II. Literature Review

A. Information Searching Behavior

As described by [1], information searching strategies can be classified into three categories: 1) search engine strategies, 2) browsing strategies, and 3) direct access strategies. The information searching behavior is highly dependent on users' goals, when exploring keyword search techniques, people's intentions should also be considered. According to [4], people's intentions with regards to information searching behavior on search engines can be classified into three main categories. The first is navigational search - people already know which web sites to visit but are searching for the accurate URL links. Second, the information search focuses on searching for particular information or objects and will require further searching on returned web pages. The third category resource search entails finding assistant resources rather than information on the Internet.

B. The URL Page Ranks of Keyword Search

Previous studies on text mining rely highly on retrieved words and phrases [5]. This however, is not directly applicable to web page searching due to several reasons. First, there are no strict standards regulating web page design. Second, Search Engine Optimization techniques have been massively applied so that the contents of web pages have been modified to raise the URL page ranks in search results. Furthermore, the keyword searching approach on search engines might be too simple to represent people's on-line search intentions. Previous studies have therefore proposed methods to combine the browsing logs stored on the side of the servers with network packets to improve search results. [6]

As social network sites have become very popular recently, people can easily interact and maintain good social

relationships with friends. [7] indicate that the social annotations newly provided through the Internet are interesting and very useful. Nowadays, there are many web sites, such as del.icio.us, that provides platforms for sharing the web pages one likes. Some researchers strive to use such information to improve the results of keyword searching. [8] have proposed methods for clustering all tags by their semantic meanings as well as for recognizing representative tags of related groups. As a result, personalized product recommendations could be made. Such approach could be extended to re-arrange the page ranking of keyword search results.

To the best of our knowledge, there is no existing work combing the web logs of keyword searching and the concepts of social groups with previous techniques to build a similarity calculation model to improve the page ranks of keyword searches.

III. The Proposed Method

A. Research Framework of the Proposed Method

The study entailed first retrieving the preferences of keyword usage from web log data. As the number of keywords searched is very high, they were first categorized according to the existing taxonomy of web sites. In our research, following [9], the directories of the Yahoo! site were adopted to classify all keywords. The keyword preferences of each user were calculated as a vector to show the preferences in different categories. Furthermore, the keyword preferences of each category were averaged by grouping individuals' preference vectors with the same preferences, with this average taken as the preferences of social groups in this study.

When a user makes a query using a keyword search, besides traditional keywords and keyword category comparisons, a search engine will compare the similarities between requested keywords and the contents of web pages. The keyword preferences of each page also need to be calculated to make such comparisons. As the differences between a user's preferences and the web pages' preference vectors are calculated, customized page ranks are attained.

The main task of search engines is to consider the similarities between requested keywords and web pages. According to the minimum risk model of keyword retrieval proposed by [10], θ_Q and θ_D are used to represent the requested keywords and the web pages, respectively. When a query q arrives, two things are considered. First, $P(\theta_Q|U)$ represents the user's,query probability distribution and $P(q|\theta_Q)$ represents the probability of the current query. S represents the resources on the Internet, $P(\theta_D|S)$ is the probability distribution of web pages across all Internet resources. Finally, $P(d|\theta_D)$ represents the probability distribution of the current page among all web pages. The relevancies of the requested keywords and web pages ($\hat{\theta}_d$) are represented as $R(d_i - q)$ which are proportional to the distance of $\hat{\theta}_q$ and $\hat{\theta}_d$ (denoted as $\Delta(\hat{\theta}_q, \hat{\theta}_d)$).

The research flow of the proposed method is shown in Fig. 1.

B. Keyword Collection and Categorization

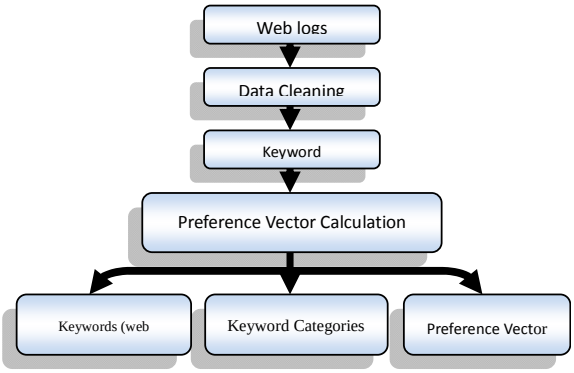


Fig. 1. The Flow of the Proposed Method.

While previous studies on keyword categorization have focused on keywords in specific application domains [11], the present study focuses on general keyword searching. The brood categories of web pages shown on the well-known portal Yahoo! site are therefore adopted to classify all requested keywords.

Let KWTree be a two-level tiers tree which stores and categorizes keywords. The first layer of the KWTree stores k categories defined by cases.

C. The Preference Vector Calculation

Let N_i be the count of keywords used by user i, $n_{i,k}$ is the frequency counts of the k^{th} categories. Let U_i be the probability distribution of keyword category usage of user i; $u_{i,j}$, the probability of the j^{th} category usage, is calculated as $\frac{n_{i,j}}{\sum_{j=1}^{k} n_{i,j}}$. U_i is the preference vector of user i. C^j be the preference vector of Category j.

We first retrieve all keywords from the users' browsing logs. Then categorize the retrieved keywords are categorized into the KWTree. In according with the KWTree, for each user i, his/her corresponding U_i is calculated. Last, the C^j of each Category j is determined.

D. The Ranks of Returned URL Pages

In this section, three main steps for ranking the returned URL pages are described. When a user issues a query in the form of a keyword search, the search engine commonly first calculates the similarities between the "requested keywords" and "the title and content of the web pages".

Let F_d be the vector to record the frequency count of keyword categories appearing in web page d. W_d is the weight preference vector of web page d.

Let P_d be the preference vector of the web page d, with $p_{d,j}$ being the preference element for the keyword category j. P_d is computed as $P_d = (w_{d,1} \times C^1 + w_{d,2} \times C^2 + \cdots + w_{d,j} \times$

$C^j + \cdots + w_{d,k} \times C^k), 1 \le j \le k$.

Once the preference vectors of all related pages are obtained, the distances between user preference and page preferences are calculated as well. The most common Euclidean distance calculation is here used as the similarity measure. Finally, all related web pages are ranked in ascending order of all distance values which are equivalent to ascending similarities. The customized page ranks are hereby created.

We first Screen related web pages by comparing requested keywords and those retrieved from web resources. Second, the weight preference vector of each related web page is calculated. Further, the web page preference vector of each related web page is calculated. Faintly, the Euclidean distance between the user and the related web page preference vector is determined. So that the customized page ranks are ordered by ascending distances.

IV. Experimental Results

The real web logs used in the present research were collected from InsightXplorer (InsightXplorer Ltd., Taiwan) which introduces on ARO (Access Rating Online) interactive database, online market research, and integrated marketing solutions. The ARO data collection provides a dynamic canvass for gathering user data in real time (at home or work). The panelists allow InsightXplorer to install the NetRover™ software to measure the overall level of internet usage and online activities. All panelists are recruited on-line and are qualified by the IX committee. As the present research focuses on keyword searching, the web logs of the top popular search engine - Google (IX Mareting Report, 2010) were used for the experiments.

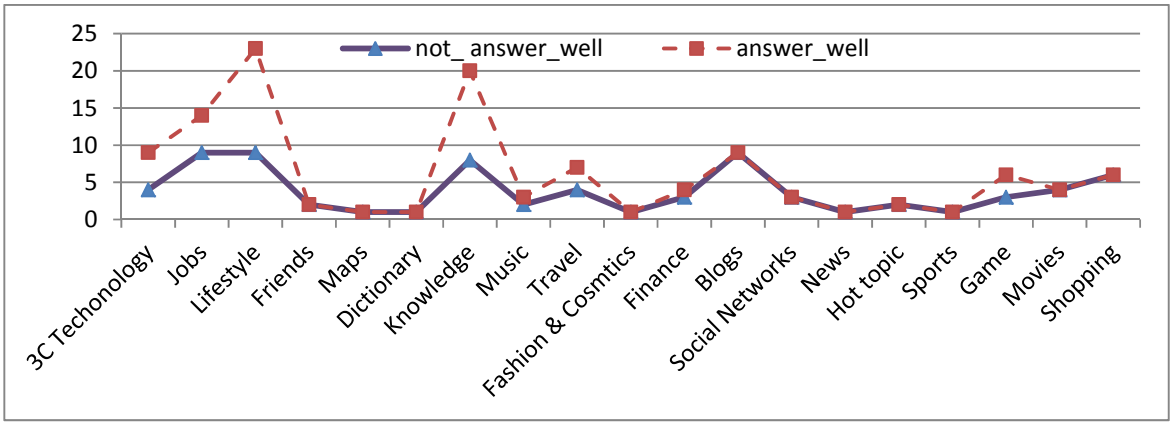


Fig. 2. The No. of Google KW Searches: not_answer_well vs. answer_well.

The first step was to classify keywords into 22 categories. Three experts with over 10 years of work experience related to the Internet with the classification. Then, the U_is of 1,122 panelists as well as the C^js of 22 categories were calculated. Thereafter, when a new query was issued, the related pages were returned via traditional keyword comparison work, the weights of related URL pages were computed, and the preference vectors of the pages were obtained. Finally, customized ranks were attained.

The Google search results as the inputs for the steps to calculate the preference vectors. According to Cacheda and Viña (2001) and surveys from SEO operators, users commonly only browse the first three pages of the returned results (the equivalent of 30 web pages). In the present study, the top 30 returned URL pages from Google search were therefore deemed to be related pages and were further used in calculating the P_ds.

The experiments done focus on the keyword searches which Google engine can recommend right hyperlinks within top 30 rankings. However during the same time period, there are some keyword searches which Google can't answer well. Fig. 2 is plotted to demonstrate the Google keyword search tend to answer better in Knowledge, Lifestyle, Jobs, 3C technology, Travel and Game than other keyword categories.

V. Conclusions

This study proposed a new method for ranking the URL links returned from search engines through keyword searches. The results show that the proposed method outperforms traditional page ranks returned by Google search. Although the method can recommend customized ranks effectively, there still has room for improvement. In the present research, only web browsing logs were collected and used to calculate the customized page ranks by considering the common interest preferences of groups. Further research could be done by utilizing real data from social network sites to make customized recommendations.

Acknowledgment

The authors would like to acknowledge the Ministry of Science and Technology, Taiwan, R.O.C. which provides supports in part under the grant NSC 102-2410 -H-128 -030-MY2.

REFERENCES

[1] R. Nachmias and A. Gilad, "Needle in a hyperstack : searching for information on the World Wide Web," *Journal of Research on Technology in Education*, vol. 34, no. 4, pp.475-486, 2001.

[2] E. Brynjolfsson, Y. Hu, and D. Simester, "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales," *Management Science*, vol. 57, no. 8, pp.1373-1386, 2011.

[3] F. Cacheda and V. Ángel, "Understanding how people use search engines: a statistical analysis for e-Business," *In the proceedings of the e-Business and e-Work conference and exhibition*, Venice, 2001, pp. 319-326.

[4] D. E. Rose, and D. Levinson, "Understanding user goals in web search", *In the proceedings of the 13th international conference on World Wide Web: ACM*, New York, 2004, pp.13-19,

[5] Y. C. Liu and C. W. Lin, "A New Method to Compose Long Unknown Chinese Keywords", *Journal of Information Science*, vol. 38, no. 4, pp.366-382, 2012.

[6] H. Huang and R. J. Kauffman, "On the design of sponsored keyword advertising slot auctions: An analysis of a generalized second-price auction approach", *Electronic Commerce Research and Applications*, vol. 10, no. 2, pp.194-202, 2011.

[7] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei and Z. Su, "Optimizing web search using social annotations", *In the proceedings of the 16th international conference on World Wide Web: ACM*, Banff, Alberta, Canada, 2007, pp.501-510.

[8] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burk. "Personalized recommendation in social tagging", *In the proceedings of the 2008 ACM conference on recommender systems*: *ACM*, Lausanne, Switzerland, 2008, pp.259-266,

[9] E. Agichtein, E. Brill & D. Susan (), "Improving web search ranking by incorporating user behavior information", *In the proceedings of the 29th annual international conference on research and development in information retrieval* : *ACM*, Seattle, Washington, USA, 2006, pp. 19-26.

[10] J. Lafferty and C. Zhai, "Document language models, query models and risk minimization for information retrieval", *In the proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval : ACM*, New Orleans, Lousiana, USA, 2001, pp111-119.

[11] J. -J. Chen, P.-W. Wang,Y.-C. Huang, and H.-C. Yen, "Applying Ontology Techniques to Develop a Medication History Search and Alert System in Department of Nuclear Medicine", *Journal of Medical Systems*, vol. 36, no. 2, pp.737-746, 2012.

Multi-Projects Scheduling Via Non-cooperative Agents Through Heterogeneous Multiprocessor Systems For Energy Efficiency

Marjan Abdeyazdan[1], Mohammad Reza Moini[2]
[1]Department of Computer Engineering, College of Electricity and Computer, Mahshahr branch, Islamic Azad University, Mahshahr, Iran.
e-mail: abdeyazdan87@yahoo.com , m.abdeyazdan@mahshahriau.ac.ir
[2]Department of Computer Engineering, College of Electricity and Computer, Mahshahr branch, Islamic Azad University, Mahshahr, Iran.
e-mail: rezamoini_it@yahoo.com

Abstract. Multiprocessor systems started a revolution in high performance computing that brought about fundamental changes in computation. This article examines scheduling of multiprojects, in which each project is assumed as an agent. Scheduling is done by assignment of agents to homogeneous and heterogeneous processors in parallel where every agent is comprised of some tasks and the related tasks in each agent constitute task graphs. Every agent has one initial point and a final point. To go from the initial point to the final point, there are some strategies, namely, various scenarios, which are selected based on the two objectives of minimizing energy consumption (computation) and minimizing the makespan. Therefore, the purpose of scheduling is to minimize energy consumption (computation) and the makespan. Our proposed solution is based on the game theory in which agents are assumed as noncooperative whereas processors are considered cooperative. Additionally, processors are of heterogeneous types. We accomplish modeling and scheduling of multiprojects via intelligent agents where each agent is a project in the model and each project or agent is comprised of several tasks. Some tasks could be done in parallel and some tasks may be repeated in different agents.

We also assume that some tasks require common resources that are accessible to all agents (projects) in multiple-agent systems. The agents are non-cooperative, and they compete with the others for the common resources. The processors are assumed cooperative and heterogeneous. All projects terminate at a final state where the joint interactions could be analyzed as a dense network game, in which we seek a stable state for tasks scheduling. For this type of networks (densed), the stable scheduling states are models of pure Nash equilibrium. In this class, we demonstrate that we will arrive at a pure Nash equilibrium in local search algorithm.

In the present article, we address the problem of scheduling and the assignment of agents to homogeneous and heterogeneous agents in multiprocessor systems. The goal is to minimize energy consumption and the makespan in view of the computational complexity, constraints, efficiency, and architectural needs.

Keywords: Intelligent Agents, Network Congestion Game, Pure Nash Equilibrium

1. Introduction

A major portion of electricity in the USA is now being consumed by computers [1] and this number is growing. A study by Dataquest [7] reported that the world-wide total power dissipation of processors in PCs was 160MW in 1992, and by 2001 it had grown to 9000MW [19]. It is now widely recognized that power-aware computing is no longer an issue confined to mobile and real-time computing environments, but is important for de1sktop and conventional computing as well. More recently, industry and researchers are eyeing multi-core processors, which can attain higher. Performance running multiple thread in parallel [16]. By integrating multiple cores on a chip, designers hope to sustain performance growth while depending less on raw circuit speed and decreasing the power requirements per unit of performance.

In this paper, we address the problem of power aware scheduling/mapping of tasks onto heterogeneous and homogeneous multi-core processor (HeMP) architectures. The objective is to minimize the energy consumption and the makespan of complex computationally intensive scientific problems, subject to the performance constraints, architectural requirements. Most energy minimization techniques are based on Dynamic Voltage Scaling (DVS). The DVS technique assigns differential voltages to each sub-task to minimize energy requirements of an application. The rest of the paper is organized as follows. In Section 2, we present some methods used to address power issues in computing systems.

Effective energy consumption requires both monitoring and scheduling of resources. In the following, we briefly describe the key related wok in the context of the proposed research. Compile time (static) techniques can be used to reduce the processor's activity[5, 7, 9, 10]. In [15], techniques for re-ordering the instructions of an application to reduce the switching activity between successive instructions are presented. This work focuses on reducing the switching activity of a data bus between the on-chip cache and main memory when instruction cache misses occur. A compilation service to reduce energy consumption of Java programs is proposed in [17]. It presents a detailed study on the relative merits and demerits of moving compilation to a server, with regards to energy. A task partitioning approach using the MIN-CUT algorithm is presented in [13]. An energy cost model is developed along with the partitioning algorithm, but the time and space complexities associated with the algorithm are very high [14]. Most task-level scheduling algorithms use utilization bound for scheduling periodic tasks [2] to maintain the timeliness of processed jobs while conserving energy. Some research has been presented in the literature for energy aware scheduling of tasks with precedence constraints [11], [19] and without precedence constrains [3] for parallel machines. We have also developed a preliminary version of energy-aware resource allocation in distributed systems for multiple tasks without precedence constraints. Our game theory based solution was shown to guarantee pareto-optimal solutions in mere O(n m log(m)) time (where n is the number of tasks and m is the number of machines in the system).

Let $A = \{A_1, \ldots, A_n\}$ be a set of n intelligent agents. We consider the problem of scheduling a set of projects $P = \{P_1, \ldots, P_n\}$. Each project $Pi \in P$ has to be done by the agent $A_i \in A$, $i = 1, \ldots, n$, and each project $Pi \in P$ consists of a set of interdependent tasks. Some tasks are executed with specialized equipment or by specialized employees. We consider such equipment or specialized personnel as common resources to be used by the agents in order to accomplish their projects. As usual, there is a limited number of employees and equipment to be used in the multi agent system. Let $T = \{t_1, \ldots, t_l\}$ be the set of different tasks to be carried out so that each agent accomplishes his project. Let $R = \{E_1, \ldots, E_k\}$ be the set of common resources to be used in the multi-agent system in order to execute the tasks. The agents are non-cooperative, and they compete with others for the use of the common resources. Scheduling is the problem of allocating limited resources to do all the tasks in a limited time of operation [4]. As it occurs in single scheduling problems, certain tasks depend on others; that is, they cannot begin until all the tasks they depend on are completed. In fact, it is possible that two different agents have to carry out the same project; $P_i = P_j$, $i = j$, $i, j = 1, \ldots, n$. Although agents doing the same project could require different quantities of products associated with that project. We can build a dependency graph or precedence graph (DAG) G_i for each project $Pi \in P$. Each task represents a node of G_i and we join the last task of the project with a special node labeled as P_i.

Each edge $(v, w) \in G_i$ meaning that the task w depends directly on the task v. Applying a topological order over each DAG G_i we can find the critical path C_i for each DAG G_i, $i = 1, \ldots, n$. The resources are common to all agents but as it happens in practical situations, the same resource is only used for one agent at a particular time; a queue of requirements for service is associated with each common resource. Furthermore, each resource $E_r \in R$ determines an increasing cost function f_r which depends on the number of agents that want to use the resource E_r. The cost of using a resource $E_r \in R$ could represent the time, the price or any other measure that an agent has to pay for using that resource. In a single scheduling problem, we consider the problem of carrying out just one project $Pi \in P$. Butin a multi-scheduling scenario where all agents compete with each other in the use of common resources, we have to analyze the global joint interaction and seek arriving at stable assignments of the scheduling of the tasks. For this analysis, we model the stable assignments as Pure Nash equilibrium of a network congestion game [2, 5].

In Section 2 we present background. In section 3 described the proposed scheduling methodology and give some concluding remarks in Section 4.

2. Background

2.1. Multi-Core Processors

Most advanced processors will surpass one billion transistors on a single chip in 2007. Although, many of the architectures support shared access of global memory or caches, effective access by a core to a memory block limits concurrent access to the same block by other cores. Memory hierarchy plays a critical role at high clock rates. Locality of data reference within a core is necessary to achieve good preprocessor performance. By optimally utilizing Authorized licensed use limited to: National Chung Cheng University. Instruction and data queues and

the functional units, these processors are capable of instruction cycle times in the range of a few nanoseconds. In the following, we briefly describe the key architectural features of multi-core machines. General purpose multi-core architectures allow multiple related tasks to be executed on different cores.

2.2. A Congestion Network for the Multi scheduling System

Let Nc be the congestion network formed by joining all final states of the DAG's Gi,i = 1, . . . , n to a final node labeled by F. And where the initial state of each Gi, i = 1, . . . , n is now an initial state of Nc. Then, Nc = ∨ni=1(Gi ∪ (Pi, F)). Usually, each project Pi ∈ P could be accomplished in different ways, that is, there could exist different paths in Gi from its initial state to its final state Pi, i = 1, . . . , n. We determine for each project Pi ∈ P the set of different paths to realize such project. So, for each agent Ai ∈ A a finite set Si = {Si,1, . . . , Si,ki} is built. Si is called the set of strategies of Ai, i = 1, . . . , n. Then, Si contains the different paths on the congestion network Nc which allow Ai to carry out the project Pi. In the sequence of tasks (a strategy) for doing a project, the order of tasks is relevant, so a different order of the tasks determines a different way of doing the same project and usually with different times. For example, S1,3 = (t1, t2, t4, t5, t7, t8, t9, t11) indicates the order of execution of the tasks to make the project P1, according to the example 1. And each task of a stategy has associated the necessary resources for attaining such task. Then, each strategy indicates a path to realize a project as well as the necessary resources to be used in that project. Regarding a single project Pi ∈ P, its minimal path Ci in Gi is now just one of the possible strategies of the agent Ai in the congestion network Nc. And the collection C = (C1, C2, . . ., Cn) of the n-minimal paths should not represent the optimum point in the scheduling of the multi-projects since the concurrent use of resources increase the cost of some (maybe all) minimal paths, in such a way that a different strategy Si,j = Ci for the agent Ai could be, in a global joint interaction, less expensive than the cost of Ci. When all agents choose one of their strategies si ∈ Si i = 1, . . . , n, a state (an action in the multiagent system) is formed e = (s1, . . . , sn) ∈ S1 X . . . X Sn. Let S = {e1, . . . , eo} be the state set of the multi-scheduling system, where each state ej, j = 1, . . . , o is a configuration of the system. Then, S = S1 X . . .X Sn and the cardinality of S is given by | S |=| S1 | ∗. . .∗ | Sn |= n1 ∗ . . . ∗ nn, where each ni, i = 1, . . . , n is the number of different paths that the agent Ai has for realizing the project Pi.

In a sequential way, given a state e = (s1, . . . , sn) ∈ S, there are n! Possible ways to ejecute all the strategies in e since any permutation of s1s2 . . . sn results in a different way to realize the n projects.

If we do not consider parallelism for performing the tasks, then the total number of possible configurations for the multi-agents system is given by n1 ∗ n2 ∗ . . . ∗ nn ∗ n!. Example 1, Let A1,A2 be two agents, A1 is responsible for filling large containers with water, while A2 is responsible for filling medium bottles. Each agent determines its strategies to accomplish its project. The corresponding DAG's; G1 and G2 are shown in figure 1 and the list of their tasks appears in table 1. The strategies S1 = {s1,1, s1,2, s1,3, s1,4, s1,5} for the agent A1, are:

s1,1 = (t1, t2, t3, t6, t7, t8, t9, t11)
s1,2 = (t1, t2, t4, t6, t7, t8, t9, t11)
s1,3 = (t1, t2, t4, t5, t7, t8, t9, t11)
s1,4 = (t1, t2, t3, t6, t5, t7, t8, t9, t11)
s1,5 = (t1, t2, t4, t6, t5, t7, t8, t9, t11)
The strategies S2 = {s2,1, s2,2} for the agent A2, are:
s2,1 = (t1, t6, t8, t9, t12, t11)
s2,2 = (t1, t8, t9, t12, t11)

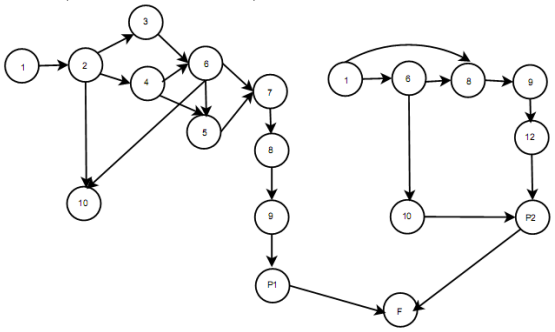


Figure 1: The congestion network of the example 1 where the nodes are the tasks.

The space of states is S = S1XS2. And for each state e ∈ S there are two possible sequences for scheduling tasks, e.g. to execute s1,j and afterwards s2,l or execute s2,l and then s1,j , j = 1, . . . , 5, l = 1, 2. Given this exponential number of possible configurations for the multi-agent system, we should apply computational methods which allow us to reduce the number of fictile solutions as well as to address the searching for optimal configurations in an adequate way. In this global scenario, there are some tasks which can be done in parallel, using different equipment or different employees. Then in a global analysis, we have to capture the maximum of tasks

which can be done in parallel. And under this consideration, the maximum number of tasks performing in parallel at the same time, coincides with the number of projects and for this reason, in this article, we are considering the same number of agents as of projects. Each task t ∈ T has associated a cost function T (t) representing the time for doing that task. Of course, if t is a single task then T (t) is just a constant time. Although, it is common that T (t) depends on the quantity of items which are manipulated through the task t. However, there are delayed times when an agent has to wait for using busy equipment or to interact with busy employees. Regarding Nc, we have to consider that some tasks require common resources and then the cost of using the resource reflects different times for t according to the number of agents in the queue waiting for using the same resource.

In any case, we assume that T (t) could be determined at any time during the scheduling system via a congestion function which depends on the number of items handled for the task t and also depends on the number of agents waiting to use the resource associated with t. We model this system as a congestion network with common resources and with a maximum of n tasks to be carried out in parallel, as well as a set of n non-cooperative agents competing among themselves for the completion of their projects. Given a state e = (s1, . . . , sn) ∈ S, si ∈ Si, i = 1, . . . , n if we consider that the sequence of tasks in e are done in parallel, maximizing at any time the number of tasks doing in parallel and if we suppose that not common resources are needed for performing the tasks, then it is not relevant the order of the execution of the n strategies and all permutations of e have equal completion time. Thus, due to the parallelism, the cardinality of the space of states is reduced to |S| = n1 ∗ n2 ∗ . . . ∗ nn, with ni = |Si|, i = 1, . . . , n. Let SPi = $\sum_{tj \in Si} T(tj)$ be the completion time for performing all the tasks involved in a strategy si ∈ Si, i = 1, . . . , n. Due to the parallelism for doing the tasks, the completion total time associated with a state e, denoted by T (e), is bounded by T (e) ≤ $\sum_{si \in e} SPi$. Furthermore, as each strategy si determines a sequence of tasks that they can not be done in parallel, then max {T (si) : si ∈ e} ≤ T (e). When the tasks require common resources in order to be performed, then there are 'delayed times' associated with the schedule of the projects since the agents in the queue of a common resource Er have to wait until Er is liberated, and then the following agent in the queue can use now the resource Er. We denote the cost as Delay (Aij) (perhaps the time) that the agent Ai has to wait for using a common resource in order to do the task tj . Given a state e = (s1, . . . , sn) ∈ S the time associated with an agent Ai, i = 1, . . . , n for such state is determined as: T (Ai, e) = $\sum_{tj \in si}$ (T (tj) + Delay(Aij)). T (Ai, e) is the time for realizing the project Pi and it is given as adding the delayed time that the agent has to wait for starting to perform each task plus the time for performing such tasks.

Note that the values Delay (Aij) i = 1, . . . , n j = 1, . . . , ni depend on the strategies of all agents in the state e and not only on the strategy choosed by Ai. Intuitively, each agent Ai can choose one strategy from among the set of strategies, but the time of the state depends on all strategies choosen by the agents. The time associated with the state e is determined as: T (e) = max {T (Ai, e) : i = 1, . . . , n}. So, the time of the state e should be the maximum completion time of the agents for performing the n projects. In order to find the equilibrium point in this scenario, we consider that each agent Ai ∈ A chooses one of his strategies si ∈ Si,i = 1, . . . , n forming a state e = (s1, . . . , sn) ∈ S. An improvement step of an agent Ai is a change of his strategy from si to s i changing to a new state e and where the time T (Ai, e) decreases with respect to T (Ai, e). Thus, we can see the neighborhood of a state e consists of those states that deviate from e only in one agent's strategy. And the improvement of the time of an agent Ai is precisely T (Ai, e) − T (Ai, e). Each state e ∈ S defines an instance of the 'job-shop' problem where individual deadlines are not considered. There are a great number of algorithms for solving this problem, see for example [4, 6, 8]. In our research, we consider a variation of the NEH heuristic [8], since it has been one of the best polynomial time algorithm for a related problem; the flow-shop problem.

3. Proposed Scheduling Approach

Since finding an optimal schedule is an NP-complete problem in general, researchers have resorted to devising a plethora of heuristics using a wide spectrum of techniques, including branch-and-bound, integer programming, searching, graph-theory, randomization, genetic algorithms, and evolutionary methods [12]. However, these methods are not applicable to energy aware scheduling where the goal is to trade-off the completion time of a parallel application with the overall energy consumption. The scheduling/mapping problem becomes a MOO problem in which the execution time is traded off with overall energy consumption. Our proposed solution is based on game theory that can solve this problem with fast turnaround time. It

also strikes a balance between the two goals. Figure 2 demonstrates the working of a hypothetical resource manager that utilizes multiple ways of achieving energy time trade-offs (These trade-off objectives are captured by multiple scenarios that are described in a later section), static and dynamic scheduling algorithms based on available resources and current state (for example temperature) of the system.

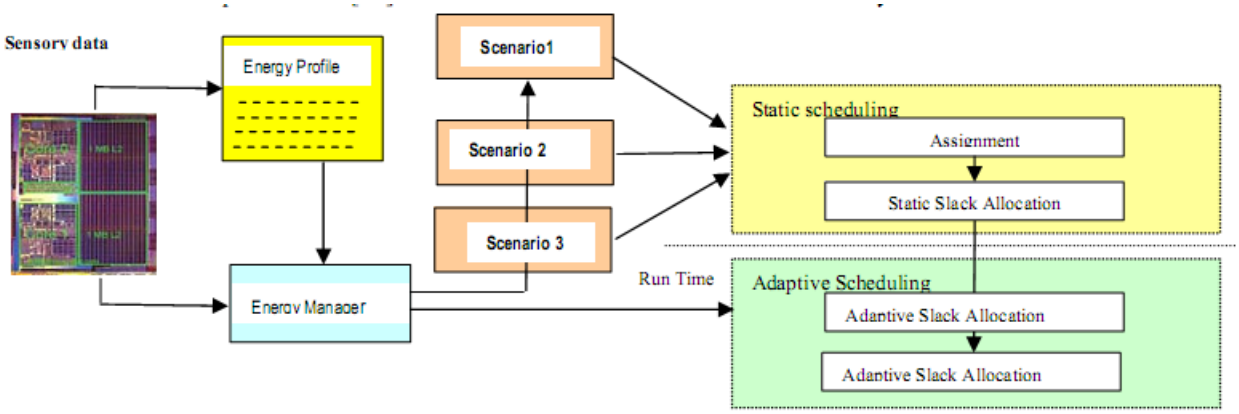


Figure 2: Energy-aware scheduling framework.

We assume an energy manager invokes our scheduling algorithms, statically and then adaptively if needed. The manager, as part of the kernel, can work under an energy profile that is calculated and updated using the sensory data for power and temperature, etc. The energy profile drives the manager to choose one of the possible scenarios that enable a scheduling algorithm to work under different objective functions and dynamically when resource conflicts occur during run-time. In the rest of this section, we briefly describe the key issues and objective functions (scenarios) that must be considered to incorporate energy-time tradeoffs by a resource manager. We then describe our prior work for with and without precedent constrained applications in separate subsections. In these subsections, we also briefly describe proposed work that is of particularly relevance to each class.

Our assumption is that a resource manager can aim for the appropriate trade-off between energy and time requirements based on current temperature conditions, priorities of the applications and pricing issues. Our goal is to provide algorithms for a rich class of energy time tradeoffs that can be effectively utilized by the resource manager. For such a multi-objectives optimization problem, there is no unique solution [8].

Figure 3 illustrates the concept of the MOO problem with conflicting objective functions. For a given application and multi-core processing machine, M, the curve AB is the pareto-optimal frontier, along which no further improvement can be done on energy consumption, P, or schedule length, SL, without sacrificing the other one. Note that points B and A are the operating points for the schedule length minimization problem and energy minimization problem, respectively. By formulating a MOO problem, we can choose an appropriate operating point along curve AB based on the current situation.

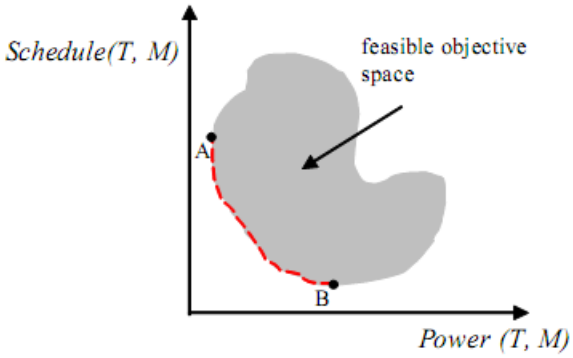


Figure 3: Illustration of MOO problem.

The remarkable property of the NBS is that it guarantees pareto-optimality and fairness. Thus, NBS provides an excellent solution to our problem because of the system environment, including the objective (collectively minimize makespan and energy consumption), the preference (pareto-optimality in terms of balancing the two objectives),

and the additional requirements (allocation is fair on all the cores in the HeMP, and hence the load is balanced). We convert the EATA problem into a cooperative game theory problem to minimize the energy consumption and the makespan simultaneously, while maintaining deadline constraints. A high complexity min-min-max optimization problem is converted using elegant cooperative game theoretical techniques into a low complexity max-max-min optimization problem. Besides lower complexity, the main benefit of this conversion is that we can always guarantee that the max-max-min optimization problem has a Bargaining Point and subsequently results in pareto-optimality and fairness. We derive an algorithm (called NBS-EATA) for obtaining NBS for the cooperative EATA game. Our NBS-EATA technique combines the classical game theoretical techniques with the Kuhn-Tucker conditions and the Lagrangian to derive a technique for identifying the Bargaining Point quickly.

Conclusion

We demonstrate a dense network game by modeling multiprojects scheduling with noncooperative agents (non-interactive). In this system, we assume that some tasks require common resources and there is also a possibility for tasks parallelism among different projects. We also show that a certain pure Nash equilibrium exists for this type of network that balances a multi-agent system in a state of e = (s1, . . . , sn). In a pure Nash equilibrium e = (s1, . . . , sn), each agent has chosen an optimal strategy S_i from i=1,2,…, in the form of a best response to the choice in the total multi-agent system.

As to the scheduling problem, we point out that cooperative game for resource recognition is better than noncooperative game (12), which is the incentive and motivation for the new method. By employing min-min-max or max-max-min method, the most suitable result could be obtained through cooperative games with less complexity. We change the processors problem to a cooperative game since processors act like players that could cooperate and compromise with one another. They could benefit from the implementation of tasks and reduce energy consumption and the makespan according to the time limit. In cooperative games, the best result could be obtained by employing min-min-max or max-max-min method with less sophistication. Additionally, we show that non-cooperative game is superior to cooperative game for agent recognition. As agents act like players that are not able to cooperate and compromise, hence each agent acts independently seeking a way to perform in parallel and to use common resources. In a cooperative game, a Nash equilibrium is employed to produce the best response. The complexity of optimism is changed to a max-max-min problem with less complexity through the technique of cooperative game theory.

References

[1]AeA (formerly American Electronics Association) Report Cybernation, www.aeanet.org.
[2]T. Abdelzaher and V. Sharma, "A Synthetic Utilization Bound for Aperiodic Tasks with Resource Requirements," Euromicro Conf. on Real Time Systems, 2003 pp. 67-75.
[3]H. Aydin, R. Melhem, D. Moss , and P. Meja-Alvarez, "Power-Aware Scheduling for Periodic Real-Time Tasks," IEEE Trans. on Computers, 53(5), May 2004, pp. 584-600.
[4] A. Garrido, M. A. Salido, F. Baber, M. A. L` opez, Heuristic Methods for Solving Job-Shop Scheduling Problems, Citeseer 2000, url: citeseer. ist. psu. edu.
[5]L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley, ScaLAPACK Users' Guide, SIAM Publications, Philadelphia, 1997.
[6] N. Melab, M. Mezmaz, E.- G. Talbi, Parallel cooperative meta-heurisitcs on the computation algrid. A case study: the bi-objective Flow-Shop problem, Elsevier Parallel Computing 32(2006), pp.643-659.
[7]Dataquest, Electronically available at: http://data1.cde.ca.gov/dataquest/
[8] M. G. Ravetti, .G. Nakamura, C. Meneses, M. Resende, G. Mateus, P. Pardalos, Hybrid Heuristics for the permutation flow shop problem, Tech.Report AT&TLabs TD-6V9MEV,2006.
[9]Environment Protection Agency, Electronically available at: http://www.epa.gov/.
[10]J. Greenberg, The Theory of Social Situations: An Alternative Game-Theoretic Approach, Cambridge University Press, Cambridge, UK, 1990.
[11]J. Kang and S. Ranka, "Dynamic Algorithms for Energy Minimization on Parallel Machines, Euromicro Intl. Conf. on Parallel," Distributedand Network-based Processing, 2008, to appear.
[12]S.U. Khan and I. Ahmad, "'Non-cooperative, Semi-cooperative, and Cooperative Games-based Grid Resource Allocation," Int'l Parallel and Distributed Processing Symposium, 2006.
[13]M. Lee, Y.S. Ryu, S. Hong, and C. Lee, "Performance Impact of Resource Conflicts On Chip Multi-Processor Servers," Applied Parallel Computing, State of the Art in Scientific Computing, Lecture Notes in Computer Science, Springer, Vol. 4699, 2007, pp. 67-75.
[14]Y.-H. Lu, T. Simunic, and G. De Micheli, "Software Controlled Power Management," in 7th Int'l Workshop on Hardware/Software Codesign, 1999, pp. 157-161.
[15]Q. F. Stout, "Minimizing Peak Energy on Mesh-connected Systems," ACM Symposium on Parallelism in Algorithms and Architectures, 2006, pp. 331-331.
[16]S. Williams, L. Oliker, R. Vuduc, K. Yelick, J. Demmel, and J. Shalf, "Optimization of Sparse Matrix-

vector Multiplication on Emerging Multicore Platforms," Int'l Conference on Supercomputing, 2007, pp. 37-46.
[17]Rudenko, P. Reiher, G.J. Popek, and G.H. Kuenning, "The Remote Processing Framework for Portable Computer Power Saving," ACM Symposium on Applied Computing, 1999, pp. 31-42.
[18]E. J. Rudkin G. L. and Loughnan, "Vortec –The Marine Energy Solution," Marine Renewable Energy Conference, 2001, pp. 67-74.
[19]M. T. Schmitz and B. M. Al-Hashimi, "Considering Power Variations of DVS Processing Elements for Energy Minimisation in Distributed Systems," Int'l Symposium on System Synthesis, 2001, pp. 250-255.

SESSION

DATA AND INFORMATION MINING + FORECASTING METHODS + SIMULATION + CROWD-SOURCING

Chair(s)

TBA

Identification of Compromised Power System State Variables

Nathan Wallace, Stainislav Ponomarev, and Travis Atkison
Deptartments of Electrical Engineering, Cyber Engineering, and Computer Science,
Louisiana Tech University, Ruston, LA, United States

Abstract— *Securing the critical infrastructure power grid is one of the biggest challenges in securing cyberspace. In this environment, control devices are spread across large geographic distances and utilize several mediums for communication. Given the required network topology of the power grid several entry points may exist that can be utilized for compromising a control network. This article explores a cyber event detection scheme based on the Grubbs' test to classify univariate values. The test is conducted only after a power system instance has been classified as containing a cyber-event. The classification of each instance is made via principal component analysis and the Hoteling's T^2 value. A Monte Carlo simulation is used to determine a set of converging power system instances and is based on the Newton-Rhapson method to solve the power flow equations of a 5 bus power system. Results indicate successful classification at a rate of 90%.*

Keywords: SCADA, PLC, control systems, state estimation, intrusion detection

1. Introduction

Perhaps one of the biggest challenges of securing cyberspace, is the ability to secure the critical infrastructure power grid. This is in part due to the inter-connective nature of the power grid and how every aspect of modern life is driven by the notion of always having power available. The power grid is composed of a meshed network of geographically distributed industrial control systems (ICS) that span large distances and utilize multiple communication mediums and protocols. Such interlacement, unbeknownst to the utility provider or independent system operator (ISO), can provide an individual or nation-state with malintent direct access to the control local area network. Once the control LAN has been breached, control decisions can be made that are outside the intended operation specifications, the most harsh being a full denial of service attack. The critical infrastructure power grid has recently seen an increase in the implementation of networked solid state devices. The key goal of such influxes is to increase the number of reporting nodes in the Wide Area Measurement System (WAMS) for the purpose of billing, state estimation, grid health, and for the efficient delivery of electricity to its consumers. However, security becomes a concern when the control decisions being implemented in the power system are based on the values being reported by the nodes in the WAMS.

In a recent effort known as Project Shine, over 7,200 control devices were found to be directly connected to the World Wide Web [1]. These startling results indicate that critical control devices have and will continue to be accidentally connected in a manner that is inconsistent with the so called 'air-gap' separation. Other possible and, in some cases, historically documented breaches into power systems are conducted via insider threat, the use of a zero-day attacks, or unpatched system attacks. The approach presented in this article aims at solving the detection of attacks against power systems using a context specific approach.

The approach presented in this article uses the Grubbs' Test to identify the reporting power system node that was compromised. This analysis is made possible by first using a transformation that identifies if an instance contains a cyber-event. Specifically, principal component analysis is used as the approach for transforming power system instances, and the Hotelling T^2 metric is used for the classification of each newly observed instance. Once an instance is labeled as suspect, the state parameters contained within that instance are compared against the variances of previously observed or trusted state parameters using the normalized residual test, Grubbs' Test, in an effort that identifies the node or control device that was the target of the intrusion. The identification scheme is applied to the data resulting from a Monte-Carlo simulation using an iterative solution to the power flow equations. The iterative solution used for the development of system data is the Newton-Rhapson method and is known to be the most common approach for solving the power flow equations [2].

Details and model assumptions of the power grid are described in Section 2 and Section 3. Using this information a trusted model is created for the development and prototyping of the proposed cyber-event detection scheme. An overview of the statistical metrics used including residual test, Grubbs' test, and the dimensional transformation technique, principal component analysis (PCA), is given in Section 4. The cyber-event model outlined in Section 3.2 describes how the instances are created such that they represent a possible malicious attack on the power system or a failed sensor. Lastly, the results of the cyber-event detection scheme are presented in Section 5 followed by future work and conclusions.

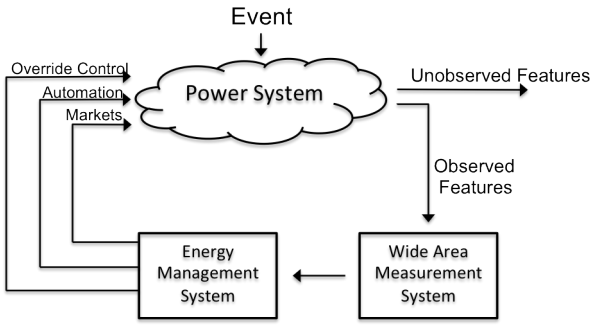


Fig. 1: Basic Power System Application Feedback Model

2. The Power Grid

A basic model, derived from similar ones presented in [3], [4], of a power system application with state feedback is presented in Figure 1. The feedback model shown is governed by the energy management system (EMS) which during an event will instruct the SCADA system to send control commands to the power system application [5]. An event can consist of a fault, i.e. a down power line, or a disturbance as modest as a customer turning on a lamp. Events change the operating conditions of the application and, if drastic enough, will cause the EMS to take immediate action to protect the system from catastrophic failure. In instances where the event does not cause immediate harm to the power system the EMS will remain idle or change control parameters to more economically provide power to customers. The purpose of this feedback interface is for constant monitoring and control of the power system application in an effort to ensure the constant and stable generation and delivery of power.

The primary steady state algorithms that determine the stability and reliability of the *critical infrastructure power grid* are: 1) Power Flow, 2) Optimal Power Flow, and 3) State Estimation. The power system challenge is to try to solve the nonlinear power balance equations in near real time given a percent of system values. The system state uses Kirchoff's Law at each power system bus throughout the system in question. Kirchoff's Law states that the sum of the powers entering a bus must be zero. The active and reactive components of the power flow equations in polar representation form from bus i to bus j can be determine by solving Equations 1 and 2.

$$0 = \Delta P_i = P_i^{injec} - V_i \sum_{i=1}^{n} V_j Y_{ij} cos(\theta_i - \theta_j - \varphi_{ij}) \quad (1)$$

$$0 = \Delta Q_i = Q_i^{injec} - V_i \sum_{i=1}^{n} V_j Y_{ij} sin(\theta_i - \theta_j - \varphi_{ij}) \quad (2)$$

where, P_i^{injec} and Q_i^{injec} are the injected powers into each bus, V_i is the voltage on bus i and Y_{ij} is element ij of the admittance matrix. Optimal power flow is the result of finding the desired power system state variables based on one or multiple cost functions. Examples of cost functions include minimization on power losses and fuel costs of generation. State estimation describes the process of estimating the state of the power system based on an incomplete picture of the system being observed. With state estimation, system parameters are measured using intelligent electronic devices (IEDs) and are reported back to the SCADA system.

2.1 State Estimation and Power Flow

Power flow analysis uses an iterative method, in most cases the Newton-Rhapshon method [2], for solving the nonlinear algebraic power flow equations, Equations 1 and 2 [6]. Convergence is said to happen when the error or mismatch drops below a certain threshold. For instance, the error stopping point used in this approach is $\varepsilon_s = 0.01$. This means that the absolute values of both the active and reactive power mismatches all had to be below 0.01 to be considered a converging instance. Also, for this examination convergence had to occur within 15 iterations or the instance was declared a non-converging instance. On average the 5 bus systems converged within 4 iterations. The extreme of 15 iterations was selected as a stopping point given that if the system did not converge within 15 iterations it is likely for that given set of inputs the system cannot exist. The fact of non-convergence corresponds to the likelihood that the power system being observed does not exist at that given set of inputs. For a more detailed description of the iterative solutions to the power flow problem the reader is encouraged to view the following referenced text [2], [6], [7].

3. Simulation Models

3.1 Power System Model

To demonstrate the identification of cyber-events a relatively simple power system was selected. Multiple instances of this model were conducted using the Newton-Rhapson method to solve the nonlinear algebraic power flow equations. Using the 5 Bus power system [7] shown in Figure 2 a series of power flow simulations were conducted. The system shown is a 100 MVA 138 kV system with the swing Bus positioned at Bus #1 or the Slack Bus. Generators are connected at Bus #1 and Bus #2. Loads are connected to every Bus in this model and are identified by that Bus's number. Table 1 shows the impedances used for the six transmission lines considered in this system model. A snapshot of the Bus input data is shown in Table 2. This information serves as the input parameters to the power flow equations and with the successful convergence of the Newton-Rhapson method the other variables can be determined. Bus #3 is a voltage controlled Bus and is part of the input variable set. The slack Bus is simulated in such a way that given the inputs shown in Figure 2 it picks up the remaining slack to supply the required load.

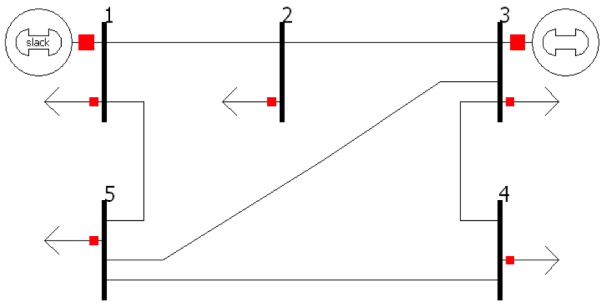


Fig. 2: Five Bus One Line Diagram [7]

Table 1: 5 Bus Transmission Line Parameters [7]

Bus - Bus	Line Length (mi)	R	X	B
1 - 2	40	0.042	0.168	0.041
2 - 5	30	0.031	0.126	0.031
2 - 3	30	0.031	0.126	0.031
3 - 4	80	0.084	0.336	0.082
3 - 5	50	0.053	0.210	0.051
4 - 5	60	0.063	0.252	0.061

3.2 Cyber-Event Model

The cyber-event model used for this detection approach is two-fold in that it represents two possibilities that can occur in a power system. Event #1 can be considered to be a non-malicious incident in which the controller or sensor in the field making the measurement breaks or becomes damaged as a result of natural causes. Some examples of this may include natural disasters, faulty equipment, or wear on the device over the years. Event #2 can be classified as an actual malicious event in which an attacker purposely launches an attack against the control system. Examples of this include the falsification or spoofing of data values reported from a smart meter as revealed by Brinkhaus et al [8]. This work currently makes no distinction of the two events only that it is able to determine that an event occurred. Once detection has occurred that instance then can be further investigated and the actual cause of the event can be determined.

The approach presented in this article assumes that both Event#1 and Event#2 will produce a state value of zero at the origin of the event. This assumption provides an initial starting point for the development of the detection scheme presented in this article. Furthermore the cyber-event model assumes that only one cyber-event occurs per instance and hence forth makes no distinction between the two events based on the developed identification scheme. An alarmed instance will only show that either event could have been the cause of the cyber-event.

To simulate these types of events a random instance from data matrix $\mathbf{X}$ was selected. This random instance vector $\overrightarrow{X_r}$ serves as the basis for the event simulation. Currently a total of ten events are simulated each event corresponds to an instance or row in a new suspicious data set $\mathbf{X}'$. For the first event, first row in the suspicious set, the variable x_1 of $\overrightarrow{X_r}$ is changed to a zero representing either a failure or an attack occurring at the voltage reading on Bus #1. This is done while holding all other values equal to the corresponding variables of the random vector $\overrightarrow{X_r}$. For each subsequent event instance the next variable is changed while holding all variables consistent with the values from $\overrightarrow{X_r}$. All simulated events occur at the bus voltages and the real power at each of the 5 loads.

Table 2: 5 Bus Input Snapshot

Bus #	Type	V	Delta	PG	QG	PL	QL
1	0	-	0	-	-	0.65	0.3
2	1	-	-	0	0	1.150	0.6
3	2	1.020	0	1.8	-	0.7	0.4
4	1	-	-	0	0	0.7	0.3
5	1	-	-	0	0	0.850	0.4

4. Identification Approach

4.1 Principal Component Analysis

In any determinable system there is a finite number of driving forces which governs how the system behaves. By observing grouping phenomenon in the data it is possible to replace a group of variables with a single new variable, greatly reducing the redundancy in the data. Principal component analysis (PCA) is a quantitative process for achieving a system simplification. A decrease in redundancy and an overall simplification of the data is made possible through a transformation into a new vector space where all the basis vectors are independent of each other. The basis vectors in the new dimensional space are called principal components [9]. PCA is based on the statistics of a training set to linearly transform the set in such a way that the new primary basis are independent of each other. The linear transformation used is based on a covariance matrix which is defined by the patterns found in the training set. PCA finds a linear transformation such that

$$\mathbf{Y} = \mathbf{W}\mathbf{X} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are mxn matrices related by a transformation $\mathbf{W}$. Based on Equation 3 the following variables can be defined: $\mathbf{w_i}$ are the rows of $\mathbf{W}$, $\mathbf{x_i}$ are the columns of $\mathbf{X}$, and $\mathbf{y_i}$ are the columns of $\mathbf{Y}$.

The row vectors of $\mathbf{W}$ $\{w_1, ..., w_m\}$ are called the principal components of $\mathbf{x}$. Before PCA can be applied to a data set it is customary to first preform sanitization on the data. This sanitization guarantees any unintended biassing of the new components. After centering the normalized covariance $\mathbf{S_X}$ was determined using the unbiased estimator for normalization.

$$\mathbf{S_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \tag{4}$$

This produced a covariance matrix with dimensions mxm with the diagonal terms representing the variances and off-diagonal terms representing the covariances of data matrix $\mathbf{X}$. The closer the off-diagonal terms are to zero the closer the variables represented by the indices of $\mathbf{S_X}$ are to being completely uncorrelated. Conversely, the higher these off-diagonal terms are the more correlated the two variables are. Also the higher the off diagonal terms are the higher the redundancy is in the data matrix $\mathbf{X}$.

The linear transformation produced by PCA selects a transformation $\mathbf{W}$ such that the principal components or basis vectors w_i produced are completely orthonormal. Orthonormality is ensured due to the fact that the dot product of each basis vector with another produces the Kronecker delta function, $w_i \cdot w_j = \delta_{ij}$. In addition to being orthonormal, the basis vectors are ordered based on the amount of variance that is being accounted for by that basis vector or principal component. This corresponds to the fact that PCA will produce a transformation matrix $\mathbf{W}$ such that the variance of data matrix $\mathbf{X}$ is mostly accounted for by principal component w_1. As hinted at in the previous section the lower the diagonal terms of the covariance matrix are the lower the redundancy is in the data. Therefore the solution to PCA seeks a covariance matrix $\mathbf{S_Y}$ such that the off-diagonal terms are zero where,

$$\mathbf{S_Y} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T \tag{5}$$

Plugging Equation 3 into Equation 5 we have

$$\mathbf{S_Y} = \frac{1}{n-1}\mathbf{W}(\mathbf{X}\mathbf{X}^T)\mathbf{W}^T \tag{6}$$

With this solution to PCA it can be shown that the principal components of data matrix $\mathbf{X}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ or are the rows of $\mathbf{W}$. Also, the i^{th} diagonal term of $\mathbf{S_Y}$ is the variance of $\mathbf{X}$ projected onto $\mathbf{p_i}$.

4.2 Classification of New Power System Instances

The Naive Bayes classifier Hotelling T^2 metric, $T^2 = n(\mathbf{X}-\mu)'\mathbf{S}^{-1}(\mathbf{X}-\mu)$, is utilized for detection and is an extension of the t-test used to determine the difference between means of two independent variables. This extension allows for a statistical measure of the multivarite distance of each instance from the center of the data set in the reduced dimensional space. The result allows for the detection of instances that occur at far distances from the data center as defined by data matrix $\mathbf{X}$. The detection approach presented in this article is a probabilistic approach in describing how likely an instance is to occur. Instances that fit to the dynamics of the data matrix $\mathbf{X}$ or control set have a high likelihood of occurring while instances that lie on the boundaries are less likely to occur. It can also be shown that the Hotelling's T^2 value follows the $\mathcal{F}$ distribution as defined by Equation 7 [10]

$$T^2 \sim \frac{(n-1)p}{(n-p)}\mathcal{F}_{p,n-p}(x) \tag{7}$$

where p is the number of principal components retained and n is the number of instances in the sample space. Because over 90% of the variance is accounted for by the first 8 principal components, a value of $p = 8$ was used. The $\mathcal{F}$ cumulative probability distribution function returns the cumulative probability of obtaining a value x for given parameters p and n. Rearranging Equation 7 we can calculate that the probability of observing at least T^2 is $P(\geq T^2) = 1 - F_{p,n-p}(z)$ where,

$$z = T^2\frac{(n-p)}{p(n-1)}$$

This allows for a probabilistic metric to determine whether or not an instance is in control. If the instance is in control then it follows the dynamics as defined by the data matrix $\mathbf{X}$. Using the maximum Hotelling T^2 value as a threshold all newly observed power system instances are classified as either suspect or non suspect. The smaller the value the closer the power system instance aligns with the dynamics of the trusted model. Then upon classification a control engineer can perform further analysis to determine the root cause of the cyber-event.

4.3 Grubbs' Test

The Grubbs' test, also known as the maximum normed residual test, is used to detect outliers in a univariate data set [11] [12]. Formally the test can be defined as a means of hypothesis testing. Using the test statistic G as defined by Equation 8 the result of the hypothesis test can either be H_0 for *no outliers in the data set* and H_a if there is exactly one outlier in the data set.

$$\mathbf{G} = \frac{max|Y_i - \bar{Y}|}{s} \tag{8}$$

With Y_i representing the measured value, $\bar{Y}$ representing the sample mean, and s representing the standard deviation of the state variable it is possible to also define the critical region for each variable. The test hypothesis H_a is true if for a given data set $Y = [y_1, y_2, \ldots, y_{N-1}, y_N]$ Equation 9 holds true; with $t_{\alpha/(2N),N-2}$ denoting the critical value of the t distribution with (N-2) degrees of freedom and a significance level of $\alpha/2N$.

$$G > \frac{(N-1)}{\sqrt{N}}\sqrt{\frac{(t_{\alpha/(2N),N-2})^2}{N-2+(t_{\alpha/(2N),N-2})^2}} \tag{9}$$

For clarity the Grubbs' test is syntactically adjusted to fit the application of detecting the compromised power system state variable. For an incoming power system instance X_i, the dimensional transformation scheme, PCA, transforms it

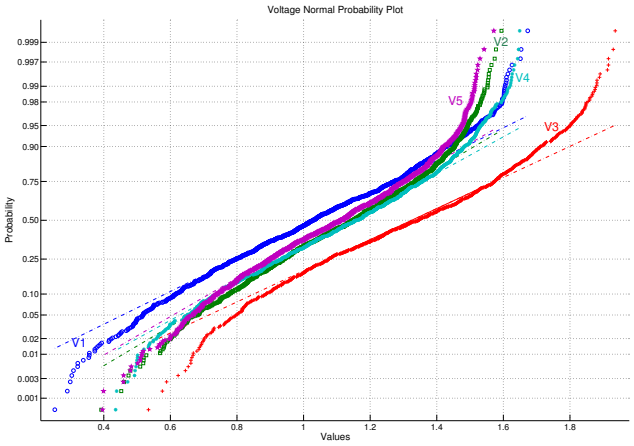


Fig. 3: Bus Voltages Distribution

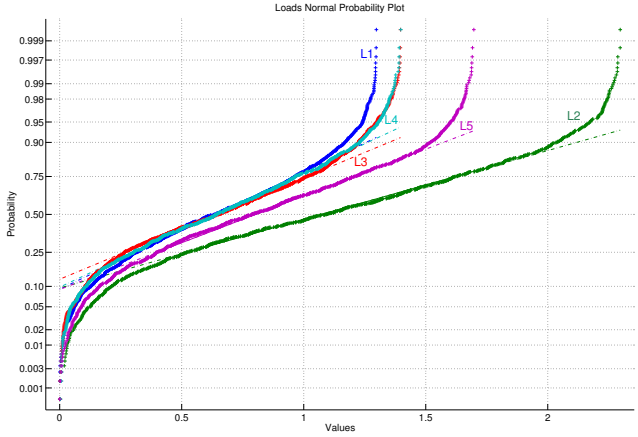


Fig. 4: Bus Loads Distribution

into a new vector space and a distance classifier is used to determine the validity of the instance. However, this does not identify the variable that was the source of the cyber-event. Therefore, after each power system instance cyber-event classification the Grubbs' test can be performed on each variable independently to determine any potential anomalies in that state variable based on historical readings. Each newly observed instance i is comprised of n variables with each variable labeled as $x_{i,j}$. By letting $Y_j = [x_{1,j}, x_{2,j}, \ldots, x_{N-1,j}, x_{N,j}]$ a new notation can be defined for the identification scheme. For instance the vector Y_1 describes the full set of bus 1 voltages.

Since the newest power system instance, if classified as containing a cyber-event, is the one under inspection with the Grubbs' test it is the N^{th} observation that will be calculated and compared. The Grubbs' test can now formally be defined as Equation 10 and 11. In Equation 11 a discriminate $\delta_{\alpha,N}$ is created that equals the right hand side of Equation 9.

$$\mathbf{G} = \frac{|Y_{N,j} - \bar{Y}_j|}{s(Y_j)} \tag{10}$$

$$\mathbf{G} > \delta_{\alpha,N} \tag{11}$$

5. Event Classification

If a cyber event has occurred it is desired to detect such an event and be able to alert on intrusion or failure. This classification capability includes the identification of the compromised node. The immediate feedback will allow the trigger of an alarm allowing a security analyst or control engineer to further investigate the event. To better understand the power system state parameters trying to be secured Figure 3 and Figure 4 show the normal probability plots for the bus voltages and bus loads respectfully for a total of 996 converging power system instances.

Given that we now have defined a transformation matrix $\mathbf{W}$ such that this transformation has eliminated all redundancy when mapped to the dimensional space we can now interpret new instances of the power system. With a trusted model derived from known instances a threshold value, T^2_{thr} was utilized to classify newly observed power system instances and is based on the maximum Hotelling T^2 of the trusted model in the transformed dimensional space. Using a trusted model containing 996 simulated power system instances, the maximum threshold value was determined to be $T^2_{thr} = 332$. When each of the 10 simulated cyber-events were mapped to the new dimensional space as a single score, that event's Hotelling T^2 value was calculated. The T^2 value calculated for each power system instance containing a cyber-event is shown in Table 3. This table reveals that the each power system instance that contained a cyber-event was successfully classified as such. However, the challenge then comes to classify the node or source of the cyber-event.

Table 3: Identification Results N=997

j	Description	T^2	$\mathbf{G}$	$\delta_{\alpha=0.05,N}$	$\delta_{\alpha=0.5,N}$
1	Bus Volt	852.31	6.266	Y	Y
2	Bus Volt	960.50	8.005	Y	Y
3	Bus Volt	909.86	7.148	Y	Y
4	Bus Volt	954.85	7.364	Y	Y
5	Bus Volt	976.88	7.850	Y	Y
6	Load 1	994.10	4.089	Y	Y
7	Load 2	995.92	3.056	N	N
8	Load 3	995.62	3.769	N	Y
9	Load 4	995.81	3.869	N	Y
10	Load 5	995.99	3.545	N	Y

Using Grubbs' test, Equations 10 and 11, a classification was conducted within each power system state variable based on the outlier hypothesis testing. Recall that the 10 simulated cyber-events correspond to malforming trusted instances by

changing only one of the variables to zero at a time. For a N value of 997 the discernment value δ for $\alpha = 0.5$ is 3.4705, $\delta_{0.5,997} = 3.4705$. Similarly the identification scheme was determined using a $\alpha = 0.5$ for N = 997, resulting in a discernment value δ of 4.039, $\delta_{0.05,997} = 4.039$. Using the calculated discriminant values combined with Equation 11 for classification, each variable can be identified as being the source of the cyber-event. The results of the identification for each discriminant value across all 10 power system state variables is shown in Table 3. A 'Y' denotes that the associated discriminant function $\delta_{\alpha,N}$ successfully identified the source of the cyber-event, and a 'N' denotes a non-successful identification.

Results indicate that for both α values, $\alpha = 0.5$ and $\alpha = 0.05$, every simulated cyber-event on the bus voltages were identified. This perhaps could have been anticipated by thoroughly analyzing Figure 3 where it is observed that a majority of the previously observed power system bus voltages occur within the region $0.6 < V_i < 1.5$. This however is not the case for the bus loads. The real power of the bus loads seem to concentrate between the region $0 < L_i < 1.2$, with a steep descent towards 0. Therefore, a cyber-event of zero on a bus load will be harder to detect than a cyber-event of zero on the bus voltages. By changing the significance value α, a larger region is covered inevitably increasing the classification potential of the discriminant classifier. However, this may lead to higher false positive identification. Using a higher alpha value, specifically $\alpha = 0.5$, all but one of the simulated cyber-events were identified.

6. Future Work

Though the simulated cyber-events offer insight into a possible detection and identification scheme based on the Grubbs' test a more complete analysis of this approach would include a full mapping of detectable regions for cyber-events of varying values. One benefit of the extensive analysis would include the fact that regions of stealthiness can be mapped out for each variable. Furthermore, future work includes a weighted alpha value that changes depending on the variance found within each power system state variable. Such a technique may decrease the false positive rate of the detection and identification scheme.

7. Conclusion

Using a normalized residual test, power system state variables were successfully identified as being the source of a cyber-event. The residual testing scheme utilized is a slight modification of the Grubbs' test to classify the newly observed power system state variables. Cyber-events are simulated by changing each power system bus voltage and the real power consumed at each bus independently to zero. A change of zero may be the result of a faulty equipment or an individual spoofing power system variables in an effort to lower his utility bill. The new observation was successfully classified as containing a cyber-event using a dimensional transformation to transform observed power system instances a probabilistic metric. Once the instance is found to contain a cyber-event the Grubbs' test was conducted to determine the power system state variable that was the source of the event. Such an analysis will allow the security investigator or control engineer to immediately isolate and fix the intrusion or problem.

The identification scheme described in this article is performed on a 5 bus power system. Trusted instances of the power system were determined using the Newton-Rhapson method of mismatch error less than 0.01 and convergence was required within 15 iterations. Principal component analysis (PCA) was used as a feature reduction method transforming 47 power system state variables into 8 principal components. Classification of each power system instance was based on a threshold Hotelling's T^2 value and if determined to contain a cyber-event the modified Grubbs' test was performed. This approach successfully classified 100% of the simulated cyber-event instances as containing a cyber-event and was able to identify 90% of the compromised power system state variables.

Acknowledgment

This research was supported by a Louisiana Board of Regents Graduate Fellowship.

References

[1] ICS-CERT, "Project shine," *ICS-CERT Newsletter Monthly Monitor*, vol. October-December, 2012.
[2] *Power Systems (The Electric Power Engineering Hbk, Second Edition)*. CRC Press, 2007.
[3] Z. Lukszo, *Securing electricity supply in the cyber age : exploring the risks of information and communication technology in tomorrow's electricity infrastructure*. Dordrecht New York: Springer, 2010.
[4] L. Grigsby, *Power system stability and control*. Boca Raton, FL: CRC Press, 2007.
[5] T. GoÌĹnen, *Electric power distribution system engineering*. New York: McGraw-Hill, 1986.
[6] J. D. Glover, M. S. Sarma, and T. Overbye, *Power System Analysis and Design, Fifth Edition*. Cengage Learning, 2011.
[7] W. D. Stevenson, *Elements of Power System Analysis (Mcgraw Hill Series in Electrical and Computer Engineering)*. Mcgraw-Hill College, 1982.
[8] C. D. Brinkhaus S., "Smart hacking for privacy," 2011.
[9] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, *Data Mining: A Knowledge Discovery Approach*. Springer, 2007.
[10] W. K. Hardle and L. Simar, *Applied Multivariate Statistical Analysis*. Springer, 2012.
[11] F. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11(1), pp. 1–21, 1969.
[12] W. Stefansky, "Rejecting outliers in factorial designs," *Technometrics*, vol. 14, pp. 469–479, 1972.

Multidimensional Scaling using Neurofuzzy System and Multivariate Analyses

Deok Hee Nam
Engineering and Computing Science, Wilberforce University, Wilberforce, OHIO, USA

Abstract - *Multidimensional scaling is one of the important techniques for a big data management. In this paper, various statistical analyses are compared to find the best-fitting method for a representation of a higher dimensional data using the reduced or smaller dimensional data using various multivariate analyses with maximum likelihood estimation through the neurofuzzy systems, which estimate the predicted output values. In addition, the estimated results are examined to find the best fitting technique through the comparison of the various statistical criteria.*

Keywords: data mining, factor analysis, maximum likelihood estimation, multidimensional scaling, neurofuzzy system, principal component analysis

1 Introduction

In these days, many scientists are interested in reducing a very large data set efficiently and equivalently without losing any significant meaning of the original data set. Among the techniques of the data reduction, the multidimensional scaling (MDS) is frequently used with applying the statistical methods such as principal component analysis, factor analysis, or clustering analysis. The purpose of multidimensional scaling (MDS) is a technique to acquire a visual representation of the pattern of similarities in order to reduce the original dimensionality to the lower dimensionality with extracting essential features by identifying the characteristic embedded components among a set of objects or data. In order to perform the feature extraction from the higher dimensional space without losing any significant meaning of the original data, statistical procedures that uses various transforming techniques (like orthogonal transformation, varimax rotation, or etc.) to convert a set of observations of possibly correlated variables (or dimensions) into a set of values of uncorrelated variables. In the paper, among different types of statistical methods, multivariate analyses including principal component analysis, factor analysis, and maximum likelihood evaluation are used to discover the newly extracted structures from the existing system without closely related measurement types. Moreover, in many cases, since the examined data system cannot be expressed by a certain mathematical expression, neurofuzzy systems are deployed to compensate the weakness of the procedures. Neurofuzzy systems can be used to compare the evaluations between the original data system and the newly reduced systems using the various statistical techniques to find out the improved solutions. In order to show the evaluation of the performance, the well-known data set, called "Air Foil Self Noise" data, is used.

2 Review of literature

2.1 Principal component analysis (PCA)

In the various statistical analysis techniques, PCA is the most popular technique along with the factor analysis to identify or extract the most meaningful components from the unknown embedded components based upon the relationship of the given variables in the original multi-variables data systems. To perform principal component analysis (PCA) [5], the total variance of the reduced or transformed data must be considered to identify the newly extracted components without losing any significant meaning of the original data and closely correlated components between the extracted components from the original variables. Ilin and Raiko [1] and Jolliffe [2] comprehensively presented all procedures of PCA. The following steps briefly introduce how to derive the steps of PCA procedures.

Let X be an n dimensional data set with $m \times n$ matrix format, i.e. $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\ldots \boldsymbol{x}_n\}$, where n is the number of measurement type to represent the dimension of the data and m is the number of samples. First, standardize X by normalizing $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\ldots \boldsymbol{x}_n$, with subtracting the mean from each measurement type. After the standardization of X, apply Singular Value Decomposition (SVD) technique to calculate the newly extracted components with the eigenvectors of the covariance. Finally, determine the dimensionality of X with most meaningful components by accumulating the calculated covariance based upon the required criterion.

2.2 Factor analysis (FA)

In factor analysis [4][12], the factors are underlying latent variables that represent the original variables. If the original variables $y_1, y_2, \ldots, y_p$ are at least moderately correlated, the basic dimensionality of the system is less than the original dimensionality, p. The purpose of using factor analysis is to reduce the redundancy among the original variables by using a smaller number of newly extracted factors.

Consider the basic structure from the vector, A, of raw data. Then, calculate the correlations (or covariance) of the response vector A, denoted by Σ by applying the basic

common factor analytic model. After the original variables are standardized, the basic input to a common factor analysis is the correlation matrix. With the correlation matrix, find the eigenvalue, λ, from the determinant equation. Using the eigenvalues, the diagonal elements of the matrix Σ are the square root of λ_i if $i = j$, and 0 if $i \neq j$. Form an initial factor loadings using Σ and Λ by multiplying each other. Then, the Varimax rotation is applied to evaluate the rotated factor loadings. Finally, applying the rotated factor loadings, 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][5][6] is a method of estimating the parameters of a statistical model which developed by R.A. Fisher in the 1920s. The probability density function (pdf), $f(y|\theta)$, for a random variable, y, and conditioned on a set of parameters, θ, identifies the observed data and 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, y, most probable, this estimation is called as maximum likelihood estimate, or MLE, of the unknown parameter, θ. Let $y_1, y_2, \ldots\ldots y_N$ be random samples from pdf, $f(y|\theta)$, where $Y = \{ y_1, y_2, \ldots\ldots y_N \}$ is the set of samples. Assuming statistical independence between the different samples, the pdf is

$$f(Y|\theta) \equiv f(y_1, y_2, \ldots\ldots y_N \mid \theta) = \prod_{k=1}^{N} f(y_k \mid \theta) \qquad (1)$$

where it is known as the likelihood function of θ with respect to Y. The maximum likelihood method estimates θ with the maximum value of the likelihood function such as

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{k=1}^{N} f(y_k \mid \theta) \qquad (2)$$

with satisfying a necessary condition to be a maximum of $\hat{\theta}_{ML}$ with respect to θ to be zero, i.e.

$$\frac{\partial \prod_{k=1}^{N} f(y_k \mid \theta)}{\partial \theta} = 0 \qquad (3)$$

Finally, the loglikelihood function can be defined as

$$L(\theta) \equiv \ln \prod_{k=1}^{N} f(y_k \mid \theta) \qquad (4)$$

and

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial \ln f(y_k \mid \theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{1}{f(y_k \mid \theta)} \frac{\partial f(y_k \mid \theta)}{\partial \theta} = 0 \qquad (5)$$

2.4 Varimax rotation

The varimax rotation was developed by Kaiser (1958) [8] 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. 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 [11] is a hybrid system which is a fuzzy system applied by neural network technique that uses a learning algorithm derived from examined and trained data to determine the developed system's characteristics. Jang [7] introduced Adaptive Neuro-Fuzzy Inference System (ANFIS), 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.

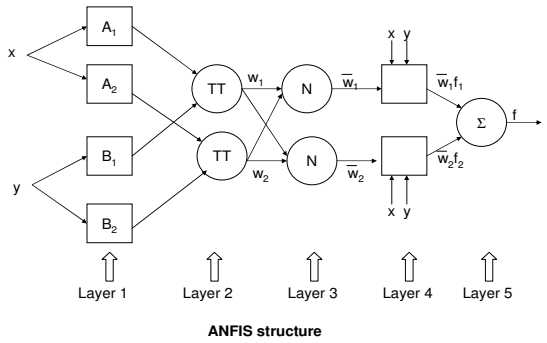


Fig. 1 Adaptive Neuro-Fuzzy Inference System (ANFIS) [7]

As shown in Fig. 1, there are five layers of ANFIS. Layer 1 consists of the fuzzy set generalized the membership functions and associated with the adaptive node with a output node. Layer 2 multiplies the incoming signals and outputs the products which represent the firing strength of the rule. In

general, this stage is called as a fuzzification of the system. Layer 3 calculates the ration of the i^{th} rule's firing strength to the sum of all rules' firing strengths called normalized firing strengths. Layer 4 is an adaptive node with a node function as consequent parameters. Layer 5 computes the overall output as the summation of all incoming signals. This is called as defuzzification of the system.

The output from Layer 5 from Fig. 1 by Jang [7] can be expressed as

$$O_{5,i} = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad (6)$$

with applying the following rulebase [2], such as

Rule 1: IF x is A_1 AND y is B_1
THEN $f_1 = p_1x + q_1y + r_1$

Rule 2: IF x is A_2 AND y is B_2
THEN $f_2 = p_2x + q_2y + r_2$.

3 Data of air foil self noise

The air foil self noise data set is obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. The NASA data set comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. For the attributes, the inputs are the frequency in Hertz, the angle of attack in degrees, the chord length in meters, the free-stream velocity in meters per second, and the suction side displacement thickness in meters. The only output is the scaled sound pressure level, in decibels.

4 Applied neurofuzzy systems

There are five different neurofuzzy systems using Adaptive-Network-Based Fuzzy Inference Systems (ANFIS) [7] to develop the procedures of estimating the suction side displacement thickness with reduced components using principal component analysis (PCA) and factor analysis with varimax rotation or maximum likelihood estimation from the five original measurements types and four reduced measurements types. The following figures are about the neurofuzzy system with the five original measurements types. Fig. 2 shows the properties of neurofuzzy system with five inputs and one output. As shown in Fig. 3, Gaussian Bell shape functions are used for the membership functions for each input and output for the neurofuzzy system. 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 air foil self noise data set that are too large to display in numerical form and for graphing functions for the mutidimensionalities of the air foil self noise data set.

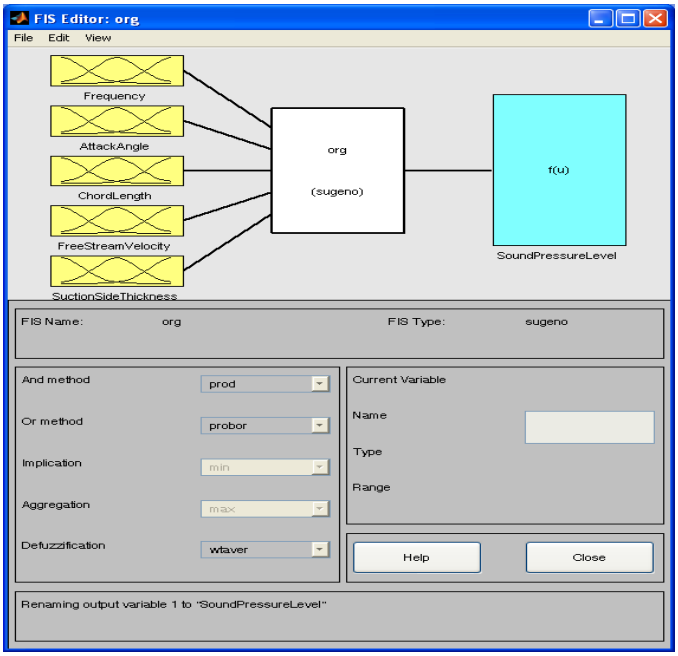


Fig. 2 Neurofuzzy inference system with properties including three inputs and an output.

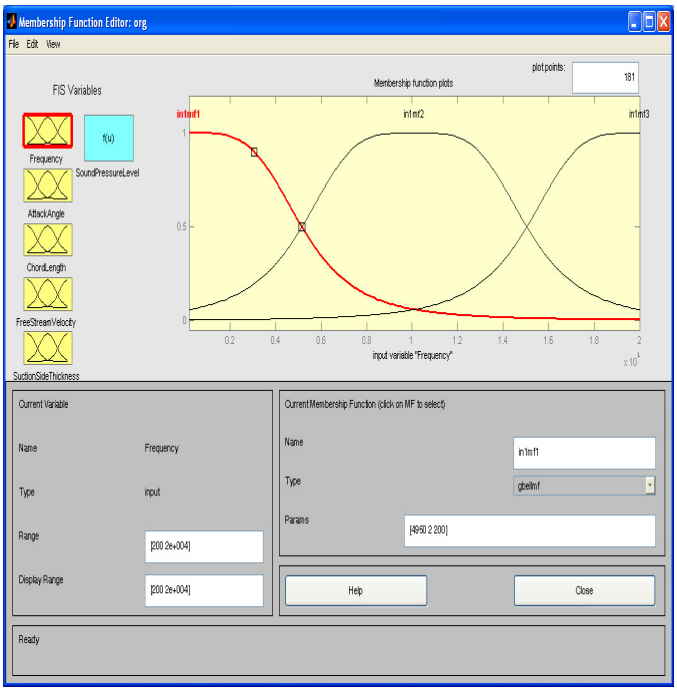


Fig. 3 Neurofuzzy inference system with membership functions

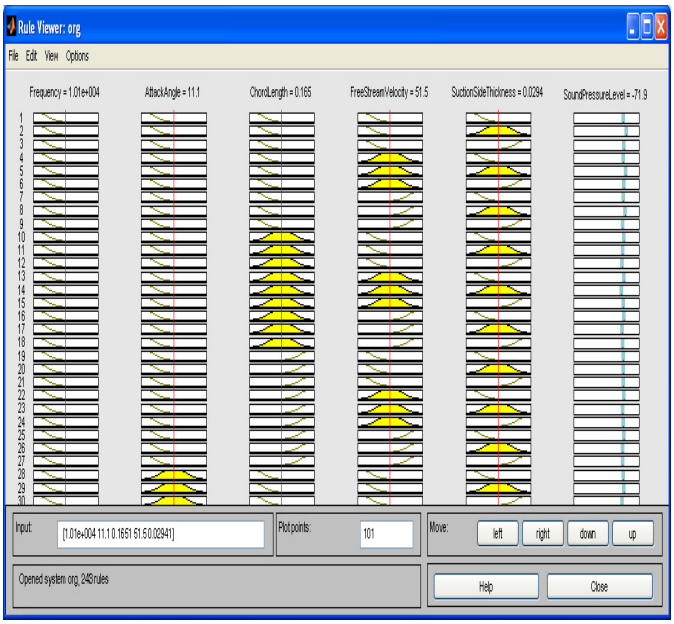


Fig. 4 Neurofuzzy inference system with applied rules for defuzzification

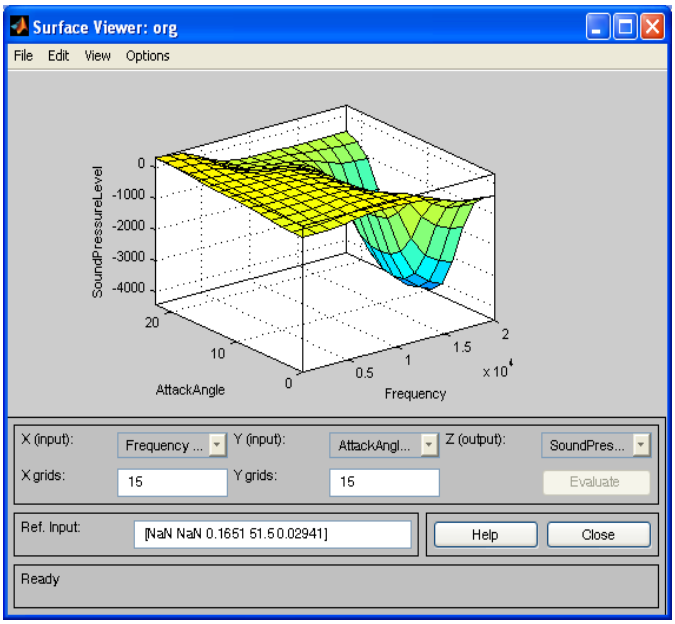


Fig. 6 Surface viewer to display in the numerical form of the air foil self noise data set

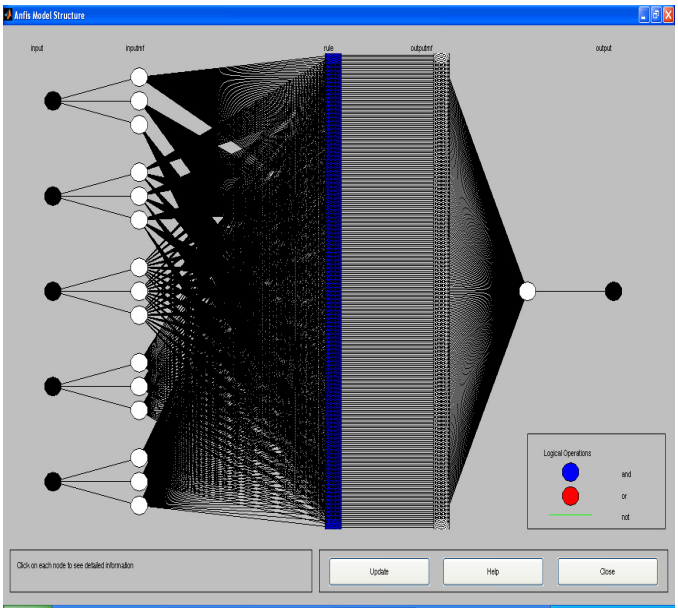


Fig. 5 ANFIS Model Sturcutre of developed Neurofuzzy inference system with three inputs

5 Analyses and results

To recognize the air foil self noise data, the scaled sound pressure level is applied as an output of each variable from the air foil self noise data. As shown in Fig. 7, all eigenvalues are presented based upon the newly transformed components from the five original measurements types. In order to decide the best-fitting reduced number of components from the five original measurements types, the accumulation of the variances from the newly extracted components using "The Eigenvalues-Greater-Than-One Rule" [9], and 0.9 or above criterion for the accumulation of the variances. There are five different techniques are compared with various neurofuzzy systems using the air foil self noise data; ORG, FM, FVM, FVP, and PCA. ORG is the neurofuzzy system using the original five components data. FM is the neurofuzzy system with factor analysis using maximum likelihood estimation. FVM is the neurofuzzy system with factor analysis using varimax rotation and maximum likelihood estimation. FVP is the neurofuzzy system with factor analysis using varimax rotation and principal components. PCA is the neurofuzzy system with principal component analysis. The reduced components are also examined with the five original measurements types using the statistical categories such as correlation (COR), root means square (RMS), standard deviations (STD), mean of absolute distance (MAD), statistical index (EWI) and error rate (ERR).

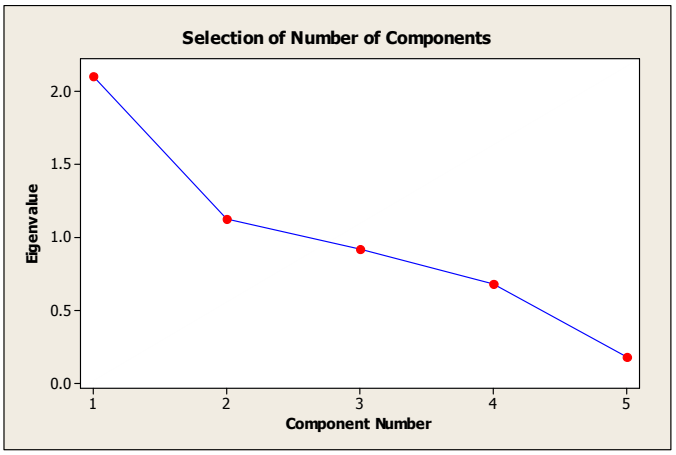


Fig. 7 The relationship between Components and Eigenvalues

TABLE 1 Statistical analysis between extracted components with estimated values and the original values using neurofuzzy systems

NFSs	COR	RMS	STD	MAD	EWI	ERR
ORG	0.8393	2.946	2.3213	2.9401	8.368	11.8531
FM	0.8771	2.4034	2.282	2.3986	7.2069	9.5877
FVM	0.8417	2.7976	2.449	2.792	8.197	11.2185
FVP	0.8347	2.8449	2.5029	2.8392	8.3523	11.4175
PCA	0.8774	2.5171	2.1378	2.5121	7.2896	10.0722

From TABLE 1, only four newly extracted components are applied to estimate the scaled sound pressure level using four inputs neurofuzzy system. In the category of the correlation, FVP shows the best performance for the evaluation. For the root mean square, FM evaluates the best matches. In the standard deviation, PCA's evaluation draws the best performance. In the category of mean of absolute distance, FM evaluates the closest pressure levels. In overall, using the category of equally weight index, FVM shows the best results among the other techniques.

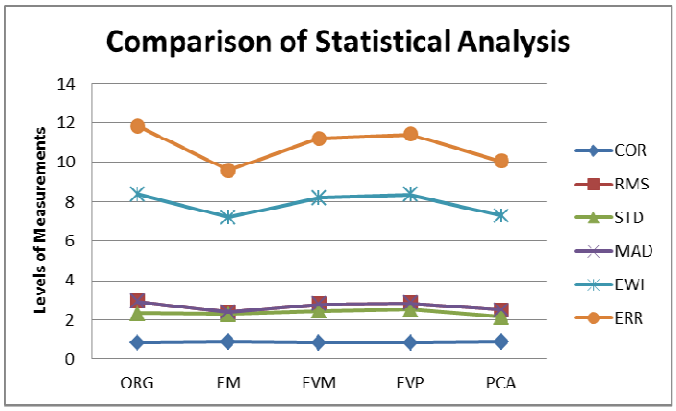


Fig. 8 Comparison of Statistical Analysis using Extracted components through neurofuzzy systems

Fig. 8 plots the statistical evaluation using five different cases with the comparison based upon the suggested statistical measurements. COR, RMS, STD, MAD, EWI, and ERR in Fig. 8 stand for the statistically evaluation with the five original components and four newly extracted components using neurofuzzy systems, respectively.

The following categories evaluate the performance of the neuro fuzzy systems using reduced data models:

Correlation (CORR): Correlation between the original output and the estimated output from the neurofuzzy system using the data from each method.

Root Mean Square (RMS): Total Root Mean Square for the distance between the original output and the estimated output using the same testing data through the neurofuzzy system.

$$RMS = \frac{\sum_{i=1}^{n}\sqrt{(x_i - y_i)^2}}{n-1} \qquad (1)$$

where x_i is the estimated value and y_i is the original output value.

Standard Deviation (STD): Standard Deviation for the distances between the original output and the estimated output using the same testing data through the neurofuzzy system.

Mean of the Absolute Distances (MAD): Mean of the absolute distances between the original output and the estimated output using the same testing data through the neurofuzzy system

Equally Weighted Index (EWI) [10]: The index value from the summation of the values with multiplying the statistical estimation value by its equally weighted potential value for each field. The value, which is close to 0, is the better results.

Error Rate (ERR): the error rate between the estimated pressure level and the original pressure level through the neurofuzzy systems generated by the examined techniques.

6 Conclusion

The presented paper describes how the original data can be efficiently identified by the less dimensional data without losing any significant meaning with evaluating the system outputs with the neurofuzzy systems. The airfoil self noise data are employed and implemented by the original data and the reduced dimensional data and evaluated by using five statistical measurements in order to compare each performance. 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 airfoil self noise data set originally adapted from the collection of the database system of UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

7 References

[1] A. Ilin and T. Raiko, "Practical Approaches to Principal Component Analysis in the Presence of Missing Values," Journal of Machine Learning Research, Vol. 11, 2010, pp. 1957-2000.

[2] I.T. Jolliffe, Principal Component Analysis, Springer, Second Edition, New York, NY, 2002.

[3] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Elsevier, Academic Press, San Diego, CA, 2003

[4] R. J. Rummel, "Understanding Factor Analysis," The Journal of Conflict Resolution, Vol. 11, No. 4, Dec. 1967, pp. 444-480.

[5] In Jae Myung, "Tutorial on maximum likelihood estimation," Journal of Mathematical Psychology, Vol. 47, 2003, pp. 90–100.

[6] M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," Journal of the Royal Statistical Society, Series B, Vol. 61, Part 3, 1999, pp.611 – 622.

[7] J.-S.R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," IEEE Trans. Systems, Man & Cybernetics, Vol. 23, 1993, pp. 665–685.

[8] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," Psychometrika Vol. 23, 1958, pp. 187–200.

[9] Norman 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, Nov. 2006, pp.151–156.

[11] Ronald Yager and Dimitar Filev, Essentials of fuzzy modeling and control, New York, John Wiley and Sons., 1994.

[12] Herve Abdi, Lynne J. Williams, and Domininique Valentin, "Multiple factor analysis: principal component analysis for multitable and multiblock data sets," WIREs Computational Statistics, Wiley Periodicals, Inc., 2013.

Mining the reviews of movie trailers on YouTube and comments on Yahoo Movies

Li-Chen Cheng* Chi Lun Huang
Department of Computer Science and Information Management, Soochow University,
Taipei, Taiwan, ROC

Abstract*—*Online reviewing is a useful and important information resource for individuals and companies. Recently several studies have focused on analyzing the reviews on Yahoo Movies where users post their comments after seeing the movies. To the best of our knowledge, there has as yet been no systematic analysis of the reviews of movie trailers on YouTube for the purpose of understanding what the consumers' feelings are. To address this challenge, we construct a framework for the summarizing and evaluating the reviews from different social media websites. Experimental evaluation shows the proposed approach has greater potential for the industry.

Keywords: opinion mining; text mining; sentiment detection

1. INTRODUCTION

The rapid growth of user-generated content on the internet has helped to make online reviewing an ever more useful and important information resource for both individuals and companies. There are several well-known web sites, such as Amazon.com and Yahoo Movies, which encourage people to post reviews detailing their experience with a product or their feelings and opinions about the movies they have watched. Recently, there have been several systematic studies of sentiment analysis and opinion mining from online review postings [1, 2]. Online product reviews have the potential to be a valuable tool for firms and manufacturers who can use them to gather feedback from their customers to further improve their products and adapt their marketing strategies [3]. Naturally, positive opinions will encourage potential consumers to adopt a product whereas negative opinions will discourage them [12]. The summarizing of customer reviews can help people to objectively evaluate their purchase decisions [4, 5, 6]. Opinion mining and summarization strategies have thus attracted increasing research attention.

In most studies the focus has been on collecting the consumer's feedback after watching movies or using the products. Several commonly used web sites, such as Amazon.com and Yahoo! Movies have designed functions to allow users to vote and rank products or movies and to filter out unhelpful reviews for readers. People tend to talk more about movies immediately after watching them and less as time goes by. Such information gathering mechanisms only work after the consumer watches a movie and then provides their feedback. Liu (2006) proved that the explanatory power of such information is somewhat dependent on the volume of online reviews. Previous studies have also found that the amount of prerelease buzz can be used as a proxy for early sales [3]. Word of mouth (WoM) also has a strong influence on people's movie selection. This study proposes a framework that will help interested firms to monitor consumer attitudes based on the analysis of reviews collected after a movie's release. The movie makers and distributors can adapt their marketing strategies and distribution tactics based on this information.

Recently, one of the most popular web sites, YouTube, has begun to provide social tools for community interaction, including the possibility of commenting on published videos and rating the comments made by other users [7]. User feedback is collected, for example from those who have watched movie trailers on YouTube which is of interest to many organizations [8]. Previous studies have also proven that online blog postings can successfully predict the ranking of book sales [9]. Gathering information on how people perceive newly released products can be helpful in the design of marketing and advertising campaigns.

Movie producers spend a lot of effort and money publicizing their movies through different mediums. This study focuses on the influence of WoM and the effect of pre-release opinions. To the best of our knowledge, there has been no study analyzing the comments on movie trailers that appear on YouTube and observing the users' behavior. This study aims to fill this gap through analysis of a large sample of text comments on Movie Trailers from YouTube.

2. RELATED WORK

The basic idea behind opinion mining is that it can be used to identify product features and to determine trends in public opinion. There are three steps in this process: (1) extracting customer comments and opinions about product features [1,2]; (2) identifying the opinion sentence in each review and deciding whether each opinion sentence is positive or negative [3,4]; (3) summarizing the results [10]. Some

This research was supported in part by the Ministry of Science and Technology Taiwan (Republic of China) under grant number NSC 102-2410-H-031 -058 -MY3.

Li-Chen Cheng is associate professor at Department of Computer Science and Information Management, Soochow University, Taipei, Taiwan, ROC (corresponding author to e-mail:lijen.cheng@gmail.com).

approaches utilize linguistic methods to discover the semantic orientations of words and sentences, in order to classify the sentiment. Other approaches extract the explicit product features based on a priori algorithm. In 2004, Hu [11] carried out a pioneering work on feature-based opinion summarization. Turney (2002) and Pang (2002) applied different methods for detecting the polarity in product reviews and movie reviews respectively.

3. THE PROPOSED FRAMEWORK

This study proposes a model for the analysis of movie comments; the architecture is diagramed in Fig.1. As demonstrated below, the comments on movie trailers from YouTube and the comments from Yahoo! Movie are gathered and then enter the pre-processing module. The data preprocessing phase is comprised of two parts: Chinese Knowledge and Information Processing (CKIP) auto tagging, and extraction of the feature opinion module. We invited several experts to score the filtered opinions to build the opinion score database. In the opinion decision module, an aggregate prelease score for each movie is compiled from the comments on movie trailers on YouTube for the opinion score database. In the same way, we accumulate a score from the Yahoo! Movie comment for the opinion score database.

3.1 Collecting data

In the first step, the data for users' comments are gathered from YouTube and Yahoo! Movie. The reviews and some metadata can also be collected from each user's review document, such as the author, the 'like' value (i.e., the number of readers who like this comment), the 'dislike' value (i.e., the number of readers who dislike this comment) and so on. For each movie, we collect all of the comments and the metadata before the opening weekend as well as the income from each movie during the same time period. After watching the movie trailer on YouTube, users are asked to express their opinion as to whether they would be willing to watch the movie in the theater. This study aims to discover the effect that the comments on these movie trailers and metadata have on sales during the opening weekend.

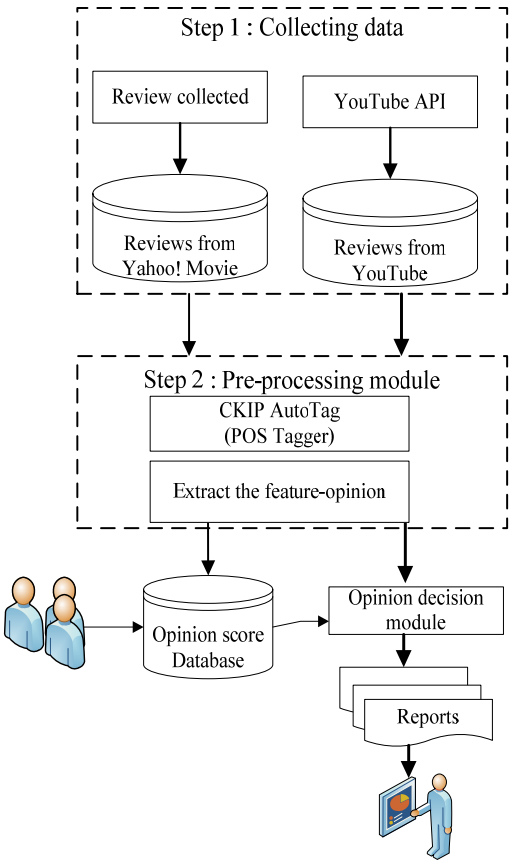


Fig. 1. An overview of the proposed framework

3.2 Pre-processing module

After the preprocessing step, the processed results for each movie are summarized to obtain a picture of the users' opinions. The steps are described as follows:

(1) CKIP Auto tag

Generally speaking, the preprocessing includes parser, part-of-speech, and feature candidate extraction functions. First, some pre-processing of words is performed including removal of stop words, stemming and so on. Next, we adopt the CKIP System to perform the Part-of-speech (POS) tagging process to produce the available datasets [24]. This process involves assigning a part-of speech (like noun, verb, pronoun, adverb, and adjective) or other lexical class marker to each word in a sentence. This process involves tokenizing every sentence into every word phrase. Each review sentence is parsed and tagged and the processing results are stored in the database.

(2) Extraction of feature opinion pairs

This step identifies the product features about which many people have expressed their opinions. Hu and Liu (2004)

established a good framework for extracting the feature opinion pairs. A feature-opinion pair consists of a feature and a relevant opinion. Product features are usually nouns or noun phrases in review sentences. First, we extract the frequent features that appear explicitly as nouns or noun phrases in the reviews. Next, we identify opinion words which are expressed as subjective opinions based on the frequent features. This study uses adjectives as opinion words. We also limit the opinion word extraction to those sentences that contain one or more product features, as we are only interested in customers' opinions about these movie features. Word features [8], opinion dictionaries [9], and syntactic structures [10-12] are used for opinion analysis.

3.3 Opinion decision module

Several domain experts were consulted to determine the scores for each opinion word which have been extracted from the previous steps. The scores are from 1 to 5. Score 5 means most consumers used this opinion word to express their satisfactions. Fig.2 is part of the opinion score database.

也不恐怖	2
也不夠	3
也不少	3
也不錯	4
也太可愛	5
也太好看	5
也太容易	3
也太弱	3
也太強	3
也太爛	1
也太少	3
也少	3
也去看	3
也可以看	3
也多	3

N2-PROJECT.You...be - dbo.Thesaurus

Opinion	Number
又怎樣	3
又其名其妙	3
又棒	4
又無聊	2
又超好笑	5
又感人	4
又溫馨	4
又緊張	4
又緊湊	4
又爛	2
下流	3
久久不能自己	5
也不好笑	2
也不(歪	3

Fig.2 part of the opinion score database.

The proposed algorithm considers both positive and negative opinions in the aggregation of a fair final score. After analyzing each review, the proposed algorithm can determine a final score.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We used the customer reviews of a few movies from the Yahoo Movies message board (http://movies.yahoo.com/). There are several reasons Yahoo Movies serves as a good source of movie WoM. In this study, we selected 104 movies that appeared in theaters from September 2013 to January 2014. We also gathered the comments on trailers for those 104 movies from YouTube. The number of comments obtained from YouTube was 4859 and the number gathered from Yahoo Movies was 3564. There were four movie genres included: action, comedy, drama and thrillers. From the algorithm, one aggregate score for each movie was obtained from comments on movie trailers from YouTube and another was determined from reviews on Yahoo Movie. The comments for five movies for each genre were summarized; the results are illustrated in Figures 3-8.

Most of the aggregated scores from the YouTube comments seem to be lower than the scores generated from reviews on Yahoo Movie. The aggregated scores only have a value of one as can be seen in Fig.3. For marketing purposes, the trailers always contain the most exciting scenes in order to attract people to come and see the movies. After many people watched the trailer of THOR, the expect value was very high shown in Fig.4. However, it was found that most viewers tended to use simple words to express their feeling about action movies meaning the comments were too short for the proposed algorithm to predict a score which would convey the true feeling of the users.

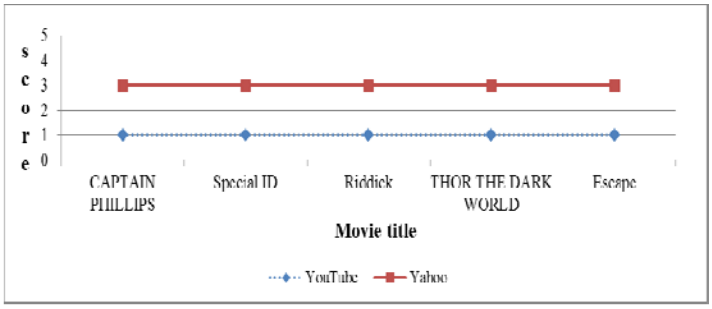


Fig.3 Results for action movies.

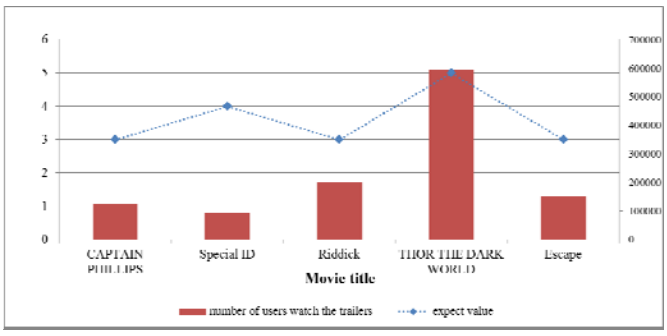


Fig.4 The number of users watched trailers of action movies and their expect values.

Examination of Fig.5 shows an interesting fact, that some of the aggregated scores based on trailers seen on YouTube are the same as the reviews from Yahoo Movie. It can be seen that, especially when the movies are popular, the users' comments contain more sentences. This makes it easier for the proposed algorithm to predict a score which is close to the thinking of users after seeing a movie. Fig. 6 is illustrated an interesting fact that the "Zone Pro Site" has a very good reputation in Taiwan. The proposed algorithm can produce a precise aggregate of the score.

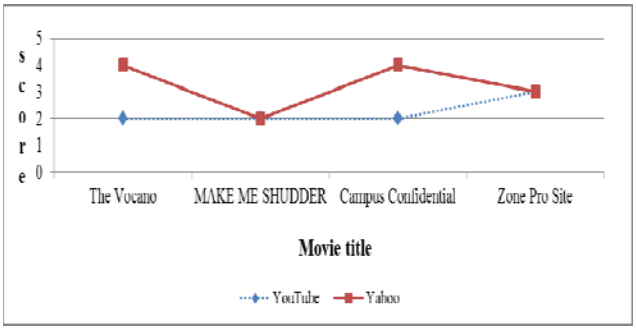


Fig.5 Results for comedies.

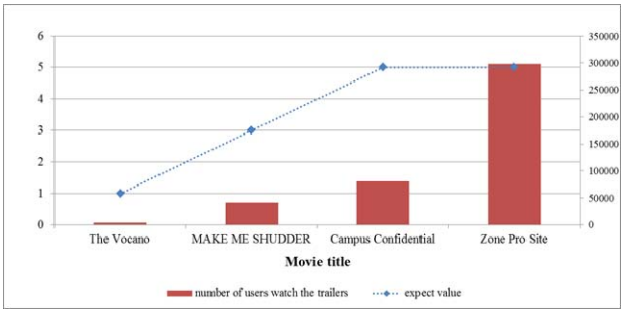


Fig.6 The number of users watched trailers of action movies and their expect values.

Two sets aggregated scores are compiled, one based on trailers from YouTube and the same for reviews from Yahoo Movie, as shown in Fig.7. It is noted that after seeing a drama, viewers may identify with the main character in the movie. They also tend to write many sentences to express their feelings in the social community. For example, "Cold Bloom" is a well-known Japanese movie featuring a story that happened after the March 11, 2011 earthquake and tsunami. The scenes are familiar enough to young people in Taiwan that they can identify with the characters. The expect value is 5 shown in Fig.8. The reviews attracted by the prelease trailer were extensive and there were more comments about their feelings left on YouTube. The richer the comments are, the more precise the score predicted by the proposed algorithm. For the genre of dramas, the users' evaluation of the trailer is the same as the feelings expressed after seeing the movie.

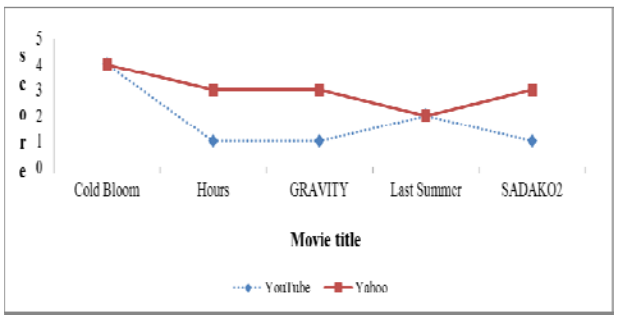


Fig.7 Results for dramas and thrillers.

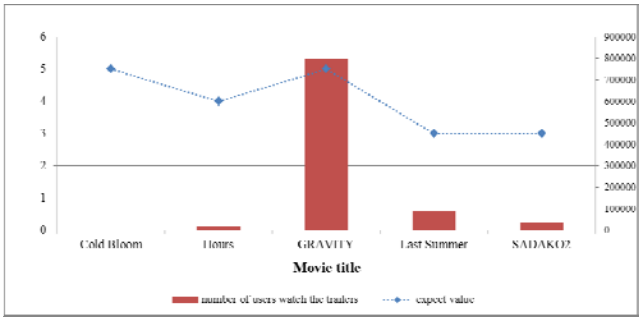


Fig.8 The number of users watched trailers of action movies and their expect values.

5. CONCLUSIONS

In this paper, a novel opinion mining framework is proposed for discovering knowledge from the reviews of movie trailers on YouTube and comments on Yahoo Movies. The objective is to automatically discover the WOM trends for online reviews of newly released products. We can observe the differences between comments on the prelease video and those made after seeing the movies. We conducted extensive experiments to evaluate the effectiveness of the proposed algorithm. We collected comments on several movies from both YouTube and Yahoo! Movie. After analyzing all the comments, we found some interesting facts about the users' behavior. The length of users' comments was influenced by the genre of the movies. In future, we will conduct more experiments to detect differences in users' comments collected from the trailer on YouTube and Yahoo movie.

ACKNOWLEDGMENT

The authors would like to acknowledge the Ministry of Science and Technology, Taiwan, R.O.C. which provides supports in part under the grant NSC number NSC 102-2410-H-031 -058 -MY3.

REFERENCES

[1] M. Hu, and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA, ACM, 2004, pp. 168-177.

[2] Q. Miao, Q. Li, et al, "AMAZING: A sentiment mining and retrieval system," Expert Systems with Applications 36(3, Part 2), 2009, pp.7192-7198.

[3] Y. Liu, "Word of mouth for movies: Its dynamics and impact on box office revenue," Journal of Marketing 70(3), 2006, pp.74-89.

[4] M. Hu, and B. Liu, "Mining opinion features in customer reviews," Proceedings of the 19th national conference on Artifical intelligence. San Jose, California, AAAI Press, 2004, pp.755-760.

[5] L.C. Cheng, Z.H. Ke, B. M. Shiue, "Detecting changes of opinion from customer reviews" In proceeding of: Eighth International Conference

on Fuzzy Systems and Knowledge Discovery (FSKD), Shanghai, China , 2011, pp.1798-1802.

[6] Siersdorfer, S., Chelaru, S., Nejd, W., & Pedro, J.S. "How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings." Proceedings of the 17th international conference on World Wide Web, 2010

[7] M. Thelwall, P. Sud, F. Vis, "Commenting onYouTube Videos: From Guatemalan", Rock to El Big Bang Journal of the American Society for Information Science and Technology, 63(3), 2012, pp.616–629.

[8] S. Choudhury, J. G. Breslin, "User Sentiment Detection: A YouTube Use Case", Proceedings of the 21st National Conference on, 2010

[9] J. A.Chevalier, D. Mayzlin, "The effect of word of mouth on sales: Online book reviews", Journal of marketing research, 43(3), 2006, pp.345-354.

[10] De Silva, I. Consumer Selection of Motion Pictures. In B. R. Litman (Ed.), The Motion Picture Mega-Industry. Needham Heights, MA: Allyn & Bacon Publishing Inc., 1998, pp.144-170

[11] Liu, Y. Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Receipts. Journal of Marketing, 70, 2006, pp.74–89.

[12] Granovetter, M., The Strength of Weak Ties. American Journal of Sociology, 78, 1973, pp.1360–1380

[13] Duan, W., Gu, B., & Whinston, A. B. (2005). "Do Online Reviews Matter? An Empirical Investigation of Panel Data", Decision Support Systems, 45, 2008, pp.1007-1016.

[14] CKIP Auto Tag. Available at: http://ckipsvr.iis.sinica.edu.tw/

[15] Peter Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics (ACL). pp. 417–424.

[16] Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

[17] J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

[18] (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

[19] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. IEEE Trans. Plasma Sci. [Online]. 21(3). pp. 876—880.

Segment, Synthesize and Repeat: CARP Paradigm for *Consumption* of Digital Content

A. Indu Anand* and **B. Anurag Wakhlu****
*Sushila Publications, P.O.Box 455, Chelmsford, Masachusetts, 01824 USA
**Coloci Inc., Chelmsford, Masachusetts, 01824 USA

Abstract. *CARP ("Computer-Aided Reading and Perusal") is a method of collecting and aggregating intelligent crowd-sourced information or data from a document or audio/video data file, which may be used to dynamically generate relevant mark-ups for documents or other consumable data files. The marked-up version of a document or data file can be displayed on demand, and may be used for purposes such as, inter alia, to enhance efficiency, comprehension and experience of reading, listening or viewing. Combined with any input technology that permits quick scanning and suitable pre-processing, CARP can use its crowd–sourced utilities to refine the highlighting and generate customized, marked up versions of the target data file for a user.*

Conf.: IKE'14; Keywords. Crowd-sourcing, reading, listening, viewing, content analysis

1 Introduction

Machine extraction of information from digital sources has emerged as a critical need, given the rate at which the corpus of digital data is accumulating. At this time, however, automated knowledge acquisition often requires human intervention to validate or improve a machine's output, on *a posteriori,* case-by-case basis, upon failure of machines. In this paper we present a crowd-sourced method of extracting information by user-specified criteria, from large files of text, audio, image or video data to assist in efficient consumption of digitally accessible content, and to enhance user experience. This approach *systematically* includes humans in the information extraction loop, and has an additional advantage: ability to dynamically provide *customized* information for a user *on demand*.

Crowdsourcing is increasingly being used for many data collection and retrieval applications, such as, online reviews of goods and services, and creation of "wikis" that allow users to collectively edit information. Computer-Aided Reading and Perusal (CARP) software discussed here extends similar collectivization of effort to the solitary activities *of actual* consumption and absorption of the information, for example, in reading, listening and viewing.

Unlike other crowd-based utilities that provide information *about* the content, the focus of CARP is to break down the content using crowd input to help with a user's *actual* consumption of the information. CARP envisions a collaboration between humans and machines, with humans in the loop from the generation to gathering to consumption of information, assisting and assisted by the machines, and thereby maintaining continual use of context to resolve ambiguities.

Further, an anticipation is *built into* CARP methodology that this loop will be used iteratively to refine the information extraction.

2 Essentials of CARP – A Crowd-Sourced Method of Information Extraction

The approach of CARP extends the idea of reading a used-book; previous readers of the book may have marked its important sections which may be used to advantage by a current reader. As the book, or document, is repeatedly read by

several users, they may continually identify further areas of the document of potential interest to other readers. Multiple mark-ups by various readers will, in general, point to the significance of sections of the book or document, in specific but *diverse* ways. This approach can aid diverse audiences achieve a panoply of objectives, e.g., to identify significant sections of a file for a specific purpose or improved comprehension by a reader of a whole document.

The essential steps of CARP are:
(1) Use crowd sourced tools to break down content into segments; (2) extract information from the segments according to the user-specified criteria, e.g., for the type, depth and manner of information extraction; and, (3) synthesize this information into a highlighted/marked-up version of the original.

The last step is a distinguishing CARP feature: Crowd-based intelligence, *in context*, is built from the experience and judgment of similar, prior users; by combining their inputs (highlights, mark-ups or comments) on demand and in accordance with the criteria specified by a present user, this method lets a present user guide information extraction, as well as the extent and manner of its display.

For the reading of a document, the "intelligent crowd" may be the previous readers of the same document who highlighted and/or provided relevant comments; for an audio or a video file it would be the prior listeners or viewers of the data file. This intelligent crowd is collectively termed "editors" or "reviewers." Notably, this group may include humans as well as machines, devices, or programs capable to provide the information sought in the segments of the data file for user-specified purposes. Using the criteria specified by a user-consumer, this approach allows for the data file to be *custom synthesized* for that consumer.

The criteria to be defined by the user(s) are envisioned to be flexible and may be collaboratively generated; a search for historical allusions in a work of fiction may be a user-specified purpose, for example, and the "relevance" of a segment for this purpose a *criterion* for highlighting it. CARP's search relies on the highlights and comments provided by previous editor-reviewers, and may be expected to capture context more appropriately with fewer spurious candidates than a typical keyword search.

The highlighting and mark up functions in CARP differ in important ways from their common usage. CARP envisions a *combination* of highlighted segments. The compilation and assembly of the segments may be done according to an algorithmic matrix, dynamically alterable in response to a user demand. For instance, if the annotations by two prior reviewers identically regard a highlighted segment as "extremely important," "important" or "unimportant" then the system will value and display the highlights their common way. But, where two reviewers disagree in classifying a highlighted section, the matrix may calculate which of the two distinct highlights to display (and how), or calculate and display a third, *synthesized* version from the two. To do this, the system calculates a "display" value {h} for *each* highlighted segment, and displays the combination accordingly.

The display value {h} for a segment may be weighted, for instance, to provide greater deference to the views of a group than of one reviewer or to those of an expert than those of a novice. Table 1 shows one simple, exemplary display value computation for "significance" of a segment, based on significance levels assigned by two users. Here 3 signifies "highly significant" and 0, "not significant," and the computed display value is used to determine how the segment is displayed - accentuating the "highly significant" and suppressing "not significant" segments.

A display of a file on a smartphone, for example, may replace a "not significant" segments with ellipsis (shown only on user demand) providing a useful, *compact* version for a small screen.

By keeping humans within the information extraction loop, CARP defines context more reliably than purely automatic methods, e.g., keyword search or statistical or filtering techniques, which are compatible with it, and which may be used additionally within the dynamic combination matrix.

Table 1

User 1 → User 2 ↓	3	2	1	0
3	3	2	2	1
2	2	2	1	1
1	2	1	1	0
0	1	1	0	0

3 CARP's Approach to Organization of Information Gathering Steps

Figure 1, reproduced from Anand, et al., [1], is a schematic of a sample CARP system that comprises a server and data storage, to which users' computers connect over a communications network, possibly the Internet or a local network. Data storage includes: Document Database, Highlight/Comment Database and a User Database. Documents may also be imported from external storage. When a user requests a document, the system also acquires the associated, integrated highlights and comments. The mark-ups may be user- or system- generated, then combined to allow for reading efficiency, comprehension, or other specified purposes. When the user submits new highlights and comments, they are combined with existing, collective highlights and comments. Generally, collectivization algorithms for highlights and comments will be different.

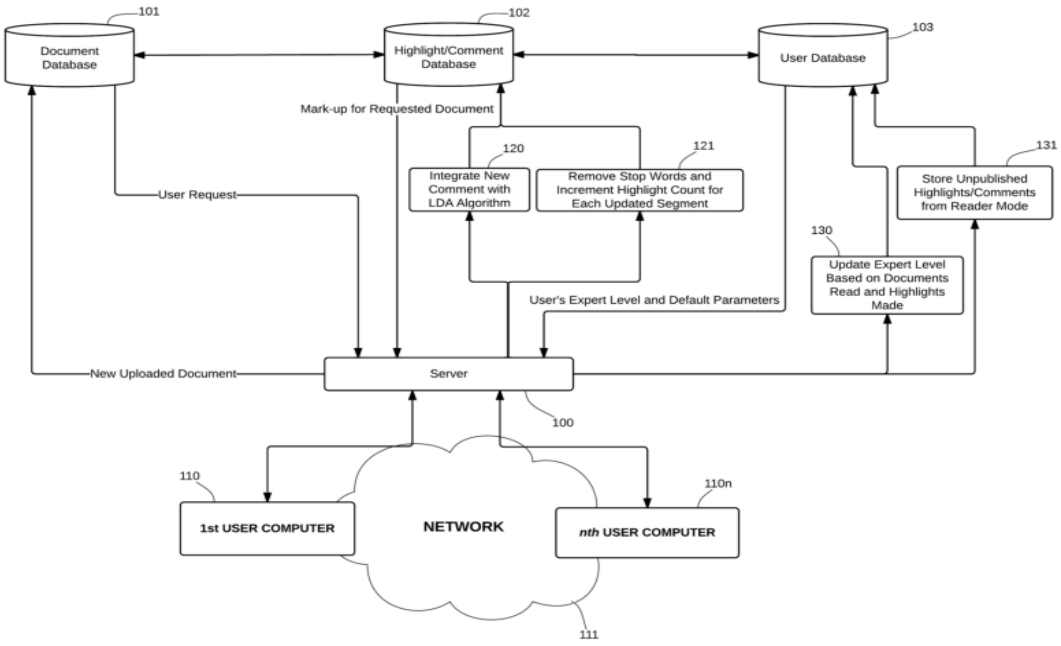


Figure 1. Schematic exemplary transaction

4 CARP's View of Information Extraction as Iteration through ***Layers*** of Information

Segments highlighting in CARP is the key to information extraction /synthesis, and it holds tangible record of the relevance or significance of a file segment. Highlighting or marking up of data files is the main functionality utilized in CARP to pass information between the system and the users.

Therefore, segment highlighting by the "editors" or "reviewers" in CARP is envisioned as a collective but *deliberative* process, resulting in a processed file that may be viewed as a collection of segments with nuanced significance levels.

Consequently, *iteration and review* become as much an expected part of the processing of a file as information collection and absorption are in learning and reinforcement.

Thus, in practice, it is useful to think of highlighting or marking-up in CARP to be a *multi-step* process. One way to implement CARP as a multi-step process could proceed as follows: In a first pass, an editor broadly identifies the pages, paragraphs or sections of significance (by suitable marking commands) for "coarse" highlighting, then returns to identify the *exact* segment(s) within the identified pages, paragraphs or sections on a second or subsequent step to refine the mark up.

For the editor/reviewer, this scheme offers major advantages. In a first pass, s/he need not slow down to highlight or comment upon the highlighted segment. Also, the editor may be tentative about the significance or objective of a highlight on a first reading but become more certain on a second or subsequent pass about its significance for a purpose or about explanatory comments, or, whether to share the highlight or keep it private, etc.

The system would save and present the first-pass, or coarse highlighting for review for possible refinement of the highlights at the start or end of a session, or on demand at any time during a session.

5 CARP for Text files

The basic commands for the mechanics of highlighting are already available in word processing packages, but their repeated use in CARP where highlighting is a key operation can be tiresome.
Essential for CARP system and user interface are the "highlighting" commands to:

- delimit the segments (for example, indicating the beginning and end of segment by a single click *in* the segment instead of repeated mouseclicks or mouse-dragging over a long segment);
- indicate differentiated *levels or tiers* of significance;
- indicate or place comments, whether or not associated with a segment highlight;
- indicate the medium (e.g., text/image/audio/video/other media) of a comment;
- indicate the *nature* of a comment (e.g., theme/subject related to the comment, or general)
- indicate "private" versus "sharable" highlighting.

The first–pass or coarse delimiters may allow the identification of words, sentences, paragraphs, or other objects such as formulas or equations for highlighting. Within a coarse segmentation differentiation may be permitted, for example, for significance level, nature of comment etc.
Typically, coarse highlighting will be in the form of a collection of *paragraphs* containing *possibly* or potentially useful information segments, which an editor/reviewer can mark appropriately with the level of significance and associated comment(s) when refining the mark-up on a subsequent pass.

A minimal, essential implementation of CARP could be straightforward: (1) find a segment of the document that is significant; (2) determine the significance level of the segment; (3) highlight the segment at the determined significance level by clicking on the appropriate "level" icon in the toolbar.

A more practical document processing scheme, however, may be the following: (1) click on the "paragraph" icon in the toolbar; (2) invoke command to indicate "level" at which one or more segments will be highlighted, by selecting the middle of the valid range, e.g., level 2 or level 3 in the range from 1-5; (3) find a segment of the document that *might* prove to be significant at *some* level; (4) click anywhere in the paragraph containing the segment to highlight the entire paragraph; (5) return to the paragraph, highlighted and saved, either during current or a subsequent session; (6) review segment previously identified as one of potential interest; (7) review the significance level of the segment; (8) raise or lower the significance level of the segment or any sub-segment; (9) raise or lower the significance level of any other segment or sub-segment within the paragraph; (10) raise or lower the significance level

of any segment in other paragraphs marked as having potential significance; (11) raise or lower the significance level of any segment in other paragraphs not marked as having potential significance during a prior pass in the process.

A very useful, novel application of CARP's approach and tools for coarse and refined highlighting is to stipulate that a file for processing may be received from *any* source and/or by any mechanism. In such cases, it is possible for an editor/reviewer, for example, to get the first-pass mark up from a program *unrelated* to CARP, and refine the highlighting by using CARP tools.

As an additional note, the coarse highlighting may be provided by a different editor/reviewer from the one refining it. In particular, it may be produced by human or non-human reviewers (e.g., an automatic summarizer/abstracter). It is possible for the first-pass highlighting to be pre-processed in CARP before the next stage, for example, by expanding an abstract and compiling the paragraphs in the original document containing the lines of the abstract. In order to provide these and similar utilities, CARP regards mark ups produced in different passes to be linked, but in a flexible manner.

6 CARP for Audio, Video, Image and Streaming Data

Coarse highlighting schemes, with appropriate modifications, extend to other digital data files such as audio, video or streamed. For example, for sound/audio in CARP we may: (i) identify audio segment by the time stamp specifying its beginning or end; (ii) specify words/text in the audio if associated with the accompanying nonverbal part of audio file; (iii) identify specific instrument from an ensemble in the audio file; and, (iv) specify a sound, note or syllable.

For a visual image or video, we may: (i) identify video segment by the time stamp specifying its beginning or end; (ii) specify words/script in the video or accompanying nonverbal part of video file; (iii) identify an object that is differentially discernable; and, (iv) identify the pixel(s) that are differentially identifiable. We may identify "coarse" level segment in this case by a tool or icon for convenient specification, for example, by drawing a box around the area of interest.

In both cases, a "coarse" segmentation may be obtained by any of these and other suitable methods, and finer segmentation may be produced in subsequent passes as in the case of text. Other remarks about multi-pass highlighting of text documents apply as well to audio and video files, including the feasibility of importation of data files, and preprocessing and processing of files within CARP utilities.

Figures 2 and 3, also reproduced from Anand et al. [1], show the progression from coarse to refined segmentation for images. A rectangular box on the astronaut's helmet in Figure 2 is the broad area of interest that would have been identified by a reviewer in a first pass. A smaller rectangular box within the first box shows the part of the picture of further interest identified in a later pass. By combining the refined segments (similar boxes) from several inputs, one may zero in on pixels of key information within the bigger box.

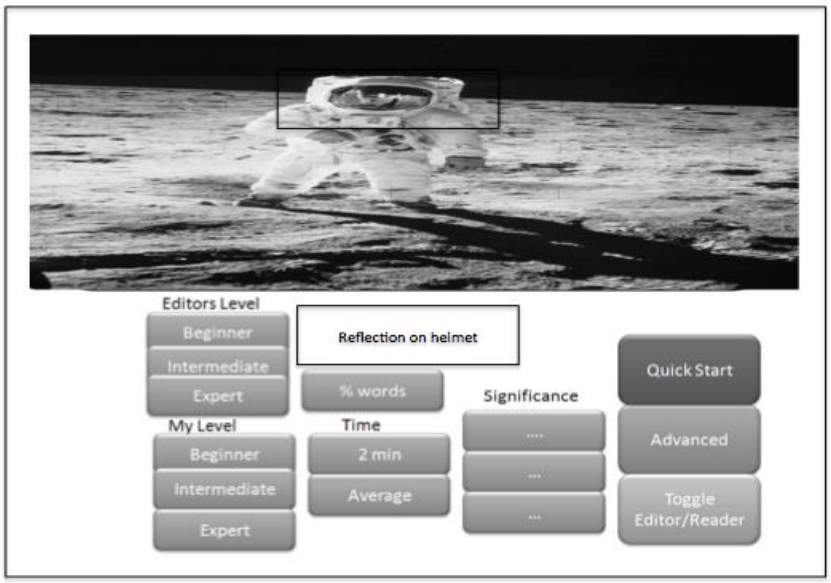


Figure 2 – Coarse Highlighting of an image: A rectangular box around the astronaut's helmet is a general area of interest.

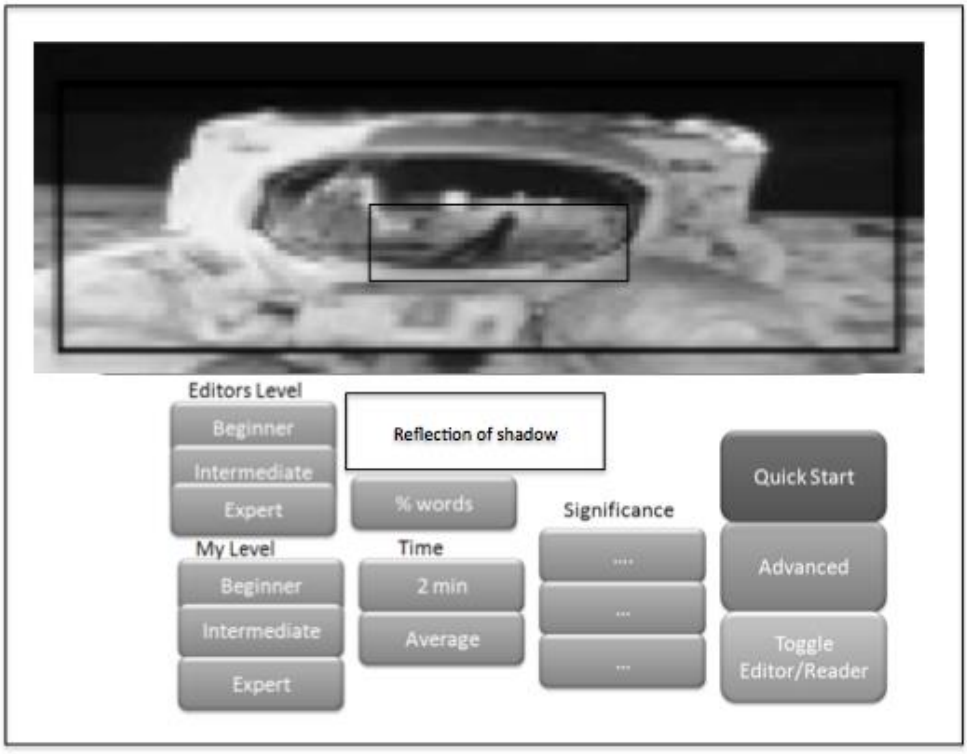


FIG. 3 -- Coarse to finer markup of the image of Fig. 2 in "multi pass" highlighting. The smaller box within the box on the helmet is the area of particular interest, and represents the "refined" highlight.

7 CARP - Conclusion

Our purpose in this paper is to show that as long as human input is indispensable to bring real world/ context information into analysis of content, it may be useful to introduce human input in a systematic way *ab initio,* and to rely on it *during* the process of content analysis.

We also show that it is possible to construct a very general approach to use human input in a flexible, adaptable manner, and yet manage to take advantage of currently available methods of information extraction in multiply-sourced, multi-layered schemes for presenting customized information for a user on demand.

Though we still need scaled studies to examine the improvement afforded by CARP's approach, the well-honed mechanism of crowd sourcing makes this approach feasible for content analysis of consumable data files.

Citations and Bibliography

[1] Anand, I.M., Wakhlu, A., Anand, P., Anand, I. : A Method And System For Computer-Aided Consumption of Information from Application Data Files. Patent Cooperation Treaty Patent Application, Publication No. WO20121625572A2 (2012)

[2] Anand, I.M., Wakhlu, A., Anand, P. : A Method of Crowd-Sourced Information Extraction From Large Data Files. To be published by Springer Verlag in the proceedings "Machine Learning and Data Mining in Pattern Recognition" of the 10th International Conference on Machine Learning and Data Mining MLDM2014.

SESSION

MODELING, SPATIAL AND SEMANTIC DATA MODELS + APPLICATIONS

Chair(s)

TBA

Structural and Percolation Models of Intelligence

Dmitry Zhukov[1], **Irina Samoylo**[2], **James William Brooks**[3] **and Victoria Hodges**[4]
[1] Professor, Head of the Department of Regional Systems of Education Quality Management, Institute of Higher Education Quality, National University of Sciences and Technology (NUST "MISiS"), Moscow, Russia
[2] Professor, Department of Medical and Biological Physics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
[3] Consultant, Salem International University, Salem, West Virginia, USA
[4] Consultant , Department of Medical and Biological Physics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Abstract *- This paper discusses the questions of the application of Percolation Theory with the purpose of estimating the number of neuronal synaptic connections sufficient for a productive intellectual activity. The use of the percolation approach in the description of human intellectual activity can be practically useful for solving problems of the creation of effective models of artificial neural networks, as well as for the development of sensitive methods for diagnostics of neural networks of the brain in autism and hyperactivity, and for the development of information security systems.*

Keywords: Modeling, mathematics, management, education, percolation

Introduction

A child is asked by his/her parents to choose a ball out of several toys. Both the parents and the child are in delight – the choice has been made correctly! Their delight is absolutely pertinent here: after all, in order for a round shaped toy to appear in the child's hand, his/her brain had to perform the most complicated intellectual work. Therefore, it is clear that discussing such a complex and multi-valued phenomenon as intelligence, it is imperative to be concerned about one of general questions of the human brain functioning and structure: the neural network granted by nature and certainly, by parents to each newborn. It was established experimentally that the number of neural connections in a cerebral cortex of a newborn is quite small. This number makes only a few percent of the neural network connections of an adult person's brain. However, in the process of child's development, the number of connections between neurons in his/her brain grows very productively and reaches its maximum by the age of six years old. In the subsequent stages of human development, there is a reduction in their neural network: the quantity of the synaptic links decreases and then it stabilizes [1]. The process of the synapses' dying off and stabilization is, probably, one of the basic mechanisms by means of which, the experience changes the structure of the brain in the course of its formation [2 – 5]. How much more reasonable is the process of reduction of the synaptic connections between neurons of the brain? What number of the synaptic connections between neurons is sufficient for a productive intellectual activity? Let us try to find the answers to these questions with the help of the methods of numerical modeling, based on the ***percolation model of human intellectual activity.***

To begin with, it is necessary to consider the basic principles of the structure and functioning of the human brain's neural network.

According to Santiago Ramon y Cajal's neuron doctrine, which is the basis for our understanding of the brain, neuron is the main structural and functional element of the brain. The dendrites of the neuron are used to obtain signals, and axons are used to transmit signals to other neurons. Signal transmission is carried out only on the special sites, the synapses, and each neuron interacts only with certain neurons. It is important to consider that in real biological structures the number of available neurons is from 10 to 100 billion, each of which has from 10 to 1000 connections with other nervous cells (**a multi connectivity condition**).

The theory of the organization of the nervous system leads to the conclusion that the brain cells, neurons, are grouped in a very complex network infrastructure, thanks to which its work is carried out. The cortex functional features are determined by the distribution of cortical nerve cells (neurons) and their connections within the layers and columns. The convergence of the impulses from various sense organs is possible. According to the modern ideas, similar convergence of diverse excitations is a neurophysiological mechanism of the integrative brain activity, i.e. the analysis and synthesis of the reciprocal activity of an organism. It is also significant that neurons are combined into complexes, which apparently realize the results of the convergence excitation in separate neurons.

Ultimately, complex interactions of all parts of the brain determine the diversity of human behavior and intellectual activity.

Abilities of the brain in information processing

John Griffith [6] made a rough calculation that if a person continually remembered information with the speed of 1 bit per second throughout 70 years of their life, then the total of 10^{14} bits would be accumulated in their memory. It is

approximately equivalent to the amount of the information stored in Encyclopedia Britannica. In fact, every second the human brain receives about 20 bits of information, and during 14 hours, it can process 18 billion bits. To store this amount of information, a person needs only one-thousandth of all nerve cells of the brain. According to various estimates, the amount of information that a person can remember for a lifetime is up to 10^{21} bits. A person is able to recall any necessary information in the tenth fractions of a second, which requires the search speed of about 50 billion bits per second. It should be noted that processing of such amount of information could be provided only by a parallel operation of the nervous structures. ***In terms of data storage and data retrieval, these structures have a very efficient topology***.

Formalized structural model of intellectual activity

A characteristic feature of the structure of the cerebral cortex is the ***oriented horizontal and vertical distributions*** of its constituent nerve cells (neurons) in the layers and columns. Thus, the cortical structure is notable for its spatially ordered arrangement of the functioning units and connections between them.

The management of intellectual processes as well as the management of motor skills can be carried out by several hierarchically subordinated rings of the connected neurons, which among themselves distribute the roles according to the hierarchy of their abilities. One part of neurons plans only general ways of realization, and the following rings of neurons are responsible for the details of the execution. However, neurons can interact not only vertically but also horizontally, forming horizontal rings. Thus, this interaction can be carried out by the principle of the branched network, where ***the trajectories of the transmission of the nerve signals look like loops***: the same signal can return to the starting point several times. The main type of the direct and reverse connections of the neo-cortex is the vertical bundles of fibers that bring the information from the subcortical structures to the cortex and send it back to the cortex. Along with the vertical links, there are intra-cortical or horizontal bundles of associative fibers extending at various levels.

The set of neurons, their communication and the topology of their connections during the process of learning and saving various images and objects, form a certain subnet of knobs. Those knobs are responsible for the process, which created them and further on, it responds for the identification of the given process (or an object).

The formalized topological model of the neural network of the brain may have the form shown in Figure 1: the arc–shaped lines represent non–overlapping connections between distant neurons, and the straight segments are the connections with the nearby nerve cells. Neurons can have both types of connections: those which are conditionally lying in one plane (one layer), and vertical connections between the nerve cells belonging to different layers (they are indicated by numbers 1, 2, 3, 4, and 5 in Fig. 1). For example, the chain of neurons indicated as **a-c-d-e-n** in Fig.1 can be included in one of the vertical layers, and the chain, which is indicated as **o-i-p-s-m-k** can be included in a horizontal layer. ***It is very important to note that the same neuron can simultaneously belong to both various horizontal and vertical layers (and the chains of the connected neurons themselves can be conditionally called horizontal and vertical rings)***.

The structure shown in Fig. 1 can be conditionally called as a random network with multiple connections between the knobs (neurons).

Let us analyze to what degree the structure represented in Fig. 1 corresponds to the existing data of the structure and operation of the human brain.

The speed of the nerve impulse is significantly lower than the speed of the transmission of electrical signals at their contacts. Therefore, practically, the brain of any person loses in speed to computers, but it considerably surpasses them in its intellectual activity because of the ***associativity of the brain's work*** and almost infinite ***degree of the parallelization of the information processing***. If there is an available network composed of 10 to 100 billion of neurons with the total number of connections of 10^{12} pieces, it can be divided into subnets (for example, the size of 10^4 – 10^6 knobs). Given that some of the neurons or their groups can be parts of different subnets, then at a rough estimate, the number of possible combinations can make:

$$C_{10^4}^{10^{12}} = \frac{10^{12}!}{10^4!\cdot(10^{12}-10^4)!} \qquad C_{10^6}^{10^{12}} = \frac{10^{12}!}{10^6!\cdot(10^{12}-10^6)!}$$

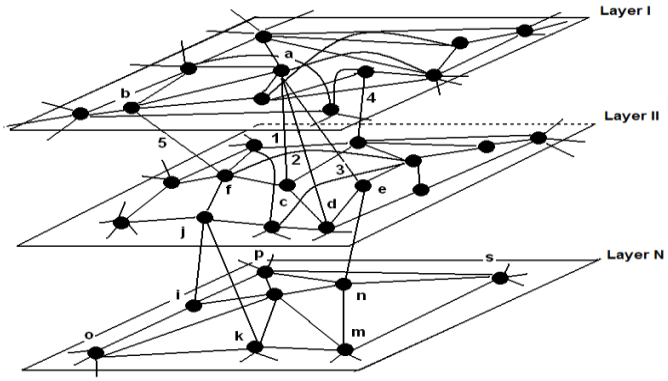


Fig. 1. Formalized model of brain topology

It is possible to consider each of such structures as a separate processor (using the Computer Science language). In this case, at a very rough estimate, there will be about 10^{11} to 10^{12} units working in parallel, which is practically not achievable (even in the distant future) for a computer built on semiconductors. For the structure shown in Fig.1, the increasing number of knobs and connections between them should lead to an increase in the number of possible subnets or (using the language of Computer Science again) parallel processors. The structure shown in Fig. 1 is not only multilayered, but ***it is also characterized by having many different connections among its knobs***. Therefore, at the reduction of its size, a removal of one part of the structure (or leaving only a small part of it) will not affect its topological properties. If the network's work is defined also by its topological properties, then its reduction in size in qualitative terms will not affect its work (the quantitative detection of

accuracy may deteriorate or time will increase). All the above is consistent with the existing data proving that the localization of functions in primary areas of the brain is duplicated in such a manner that each smallest area contains the information about the whole object.

The percolation model of intellectual activity

For image recognition purposes, we propose the following model in this paper. A signal is transmitted to the input of the neural network. The incoming signal is compared with the images previously stored in its subnet. This may result in some active (or excited) states of the neurons. If between the input and output layers of the network or in any of its sub-networks, an unbroken chain of unexploded and excited neurons appears, then the stored in the given subnet image is recognized. Otherwise, the image cannot be recognized. Then the synthesis of a certain amount of concentration of RNA and the protein coding information takes place in the neurons involved in the recognition. The above described process, creates one more subnet. The new subnet saves the new image without deleting the old ones. Considering that the number of possible created subnets is very high (about 10^{11}), it allows to store a large number of various images. Besides, given that during the recognition, the process will be taking place in parallel within all structures, then the speed of the search and access to the information should be reasonably high, despite a considerable volume of the stored data.

It is possible to name the route created through the active cells (they can be called the knobs of the subnet) as filtering or percolation. For regular structures, Percolation Theory is a well-developed area. However, the structures represented in Fig. 1 have a random irregular topology, and the study of percolation processes in them is a challenging task that can only be solved by the methods of numerical modeling.

From the point of view of a mathematician, Percolation Theory should be carried to Probability Theory on graphs. There is quite a number of monographs on both theoretical and applied aspects of percolation [7-10].

To explain the basic tenets of Percolation Theory, let us take a square network and paint over in black a part of knobs (see Fig. 2). One of the questions, which can be answered by Percolation Theory here, is: At what portion of n^c of the colored knobs does their black chain connecting the upper and lower sides of the net (the chain of connectivity), arise? For a grid of the finite size, such chains may occur at different concentration (see Fig. 2). However, if the size of grid L is directed to infinity, then the critical concentration becomes quite definite. Such critical concentration is called *the percolation threshold.*

The square grid is only one possible model. One can consider percolation on the triangular and hexagonal grids, trees, three-dimensional lattices (for example, on a cubic one that is in space with the dimension more than three). The grid does not necessarily need to be regular; it is possible to consider the processes on random lattices.

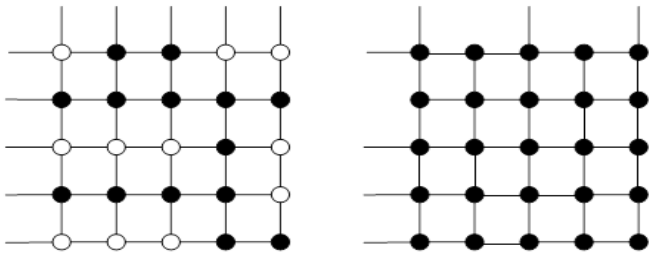


Fig. 2. Percolation on a square lattice

Let us consider percolation in random networks with multiple paths between the knobs having the form shown in Fig. 1. Let us choose any two knobs, **A** and **B,** on the opposite layers of the network and then let us start randomly activate certain other knobs. Obviously, if there are many activated knobs, a situation may occur at which between the two arbitrarily selected knobs **A** and **B,** there will be at least one "open" path (the path formed by the activated knobs). Using the methods of numerical modeling with statistical averaging of the obtained results from separate experiments, it is possible to determine at what proportion of the activated knobs (the percolation threshold) the conductivity between knobs **A** and **B** appears in the network**,** and how it depends on the average number of connections per single knob. Table 1 presents the results of numerical modeling of identifying the percolation threshold for random networks with multiple paths between the knobs (see Fig.1) with a various average number of connections per knob. (See Table 1)

Table 1.

Network type	Average of connections per knob in the network of the final size within the given structure	Portion of the activated knobs, at which conductivity occurs in the network (**n_c - percolation threshold**)
Random network with multiple paths between the knobs	2,36	0,515
	2,82	0,425
	3,29	0,365
	4,70	0,270
	4,75	0,250
	6,15	0,150
	6,17	0,185
	6,75	0,175
	9,41	0,170
	10,02	0,150
	10,31	0,130
	10,69	0,135
	11,07	0,115
	13,10	0,115

Since an increase of an average number of connections per knob in the network leads to a substantial increase in time and computing resources expenses, in numerical modelling it was decided to choose the area from 2.5 to 15 connections

per knob. Figure 3 shows the graphic dependence of the results introduced in Table 1.

Figure 3 demonstrates that at the increase in the average number of connections per knob in the network, the percolation threshold begins to pursue its certain minimum value. Thus, the received results suggest that there is no need to carry out numerical modelling for large values of the average number of connections per knob. Instead, it is possible to linearize the results and extrapolate them to higher values. (See Fig.3)

As the graphic kind of the dependence in Figure 3 reminds the exponent, then it can be described by the function of the following kind: $P(x) = P_0 e^{-z}$, where $P(x)$ – is the value of the percolation threshold with the average number of connections per knob equal to some value x, and z=1/x, P_0 – is the value of the percolation threshold at infinite number of connections per knob. As Figure 4 shows, the results presented in Table 1 are linearized well in the co-ordinates: ln P(x) – z = 1/x (natural logarithm of the percolation threshold is the reciprocal of the average number of connections x per knob), which supports the use of the function of the following kind: $P(x) = P_0 e^{-z}$. (See Fig.4)

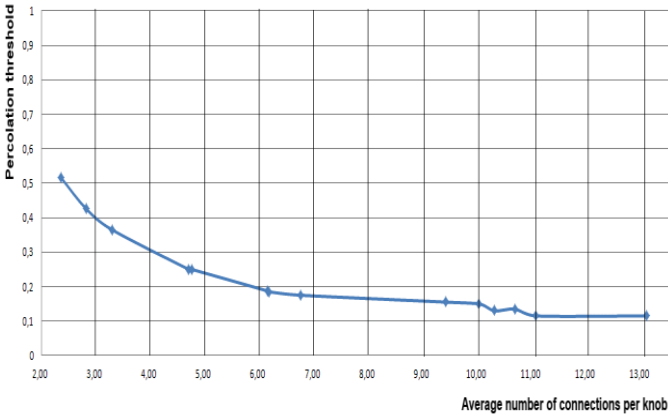


Fig. 3. Dependence of the size of the percolation threshold in a random network on average number of connections per knob

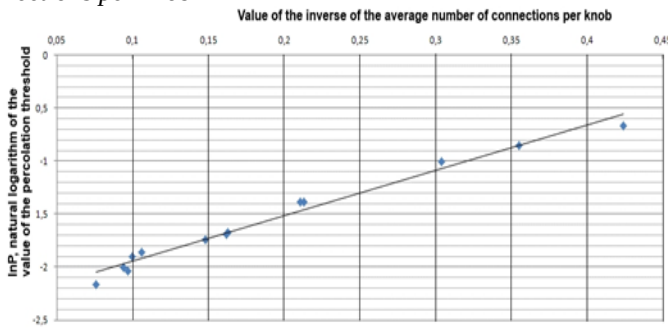


Fig. 4. Dependence of the logarithm of the value of the percolation threshold in a random network on the value of the inverse of the average number of connections per knob.

The dots in Figure 4 present the experimental data and the solid line corresponds to linear relationship:
y = 4.2882z – 2.3766, with very high correlation coefficient equal to 0.98. At z=1/x = 0 (corresponding to the case when x = ∞) we get: y=lnP_0 = –2.15, and the value of the percolation threshold at an infinite large number of connections per knob P_0 will be equal to 0.093. Thus, for a random network with an infinite large number of connections per knob, it is enough to have an equal share of the activated neurons equal to 0.093 of the total number, in order for the conductive chain of knobs to appear and for the network of knobs to recognize the presented image. With the average number of connections equal to 100, the percolation threshold is equal to 0.097, and at 10 it is equal to 0.143.

The obtained results demonstrate that a substantial increase in random network connections with an average of more than ten connections per knob, hardly changes the percolation threshold, and from the point of view of the use of the biological resources, this type of increase is unfavorable for neural structures. Therefore, the reduction of neuro-synaptic redundancy of the network is an inevitable step in the formation of a neural network of the human brain.

The use of the percolation approach can be practically helpful in solving problems in the development of more efficient models of artificial neural networks as well as in the development of sensitive methods of diagnostics of the neural networks of the brain in autism and hyperactivity. It can also be used in the design of information security systems that can detect false knobs, through which leak or substitution of information can occur.

We also consider that another potentially interesting area for the percolation model application is its use in the development of crossbar nanocomputers (while designing them a considerable redundancy of interconnections of conductive nano lattices is put in). The network reduction based on the percolation approach will allow building an optimal architecture of the keys and raising the system's resistance to defects.

Reference Literature

[1] Norman Doidge. "The Brain That Changes Itself". Appendix 2, 2010.
[2] Steven Rose. "The Making of Memory: From Molecules to Mind". New York/London/Toronto/ Sydney/Auckland: Anchor Books/Doubleday, 1993.
[3] Jean-Pierre Changeaux, Antoine Danchin. "Selective Stabilization of Developing Synapses as a Mechanism for the Specification of Neuronal Network". Nature 264, 1976, pp. 705–712.
[4] Gerald Edelman. "Neural Darwinism: The Theory of Neuronal Group Selection". Basic Books, 1987.
[5] Donald O. Hebb. "The Organization of Behavior ". Wiley, 1949.
[6] John S. Griffith. "A View of the Brain".Oxford, 1967.
[7] Dietrich Stauffer, Ammon Aharony. "Introduction to Percolation Theory". London: Tailor AND Francis, 1992.
[8] Geoffrey Grimmet. "Percolation". Berlin. Springer–Verlag, 1999.
[9] Harry Kesten, "Percolation Theory for Mathematicians". Birkhauser, 1982.
[10] Muhammad Sahimi. "Percolation Applications of Percolation Theory". London: Tailor AND Francis, 1992.

Ontology inference using spatial and trajectory domain rules

Rouaa Wannous
L3i Laboratory
University of La Rochelle, France
Email: rouaa.wannous@univ-lr.fr

Jamal Malki
and Alain Bouju
L3i Laboratory
University of La Rochelle, France

Cecile Vincent
LIENSS laboratory and UMR 7372
CNRS/University of La Rochelle, France
Email: cvincent@univ-lr.fr

***Abstract*—Capture devices give rise to a large scale spatio-temporal data describing moving object's trajectories. These devices use different technologies like global navigation satellite system (GNSS), wireless communication, radio-frequency identification (RFID), and sensors techniques. Although capture technologies differ, the captured data share common spatial and temporal features. Thus, relational database management systems (RDBMS) can be used to store and query the captured data. For this, RDBMS define spatial data types and spatial operations. Recent applications show that the solutions based on traditional data models are not sufficient to consider complex use cases that require advanced data models. A complex use case refers to data, but also to domain knowledge, to spatial reasoning or others. This article presents a sample application based on trajectories that require three types of independent data models: a domain data model, a semantic model and a spatial model. We analyze each of them and propose a modeling approach based on ontologies. This work introduces a high-level trajectory ontology and a generic spatial ontology. Also, we present our ontology matching approach for integrating the sample trajectory domain to both defined ontologies. This work has a special focus to the problem of defining ontology inference using business rules combined with spatial rules. We present an implementation framework for declarative and imperative parts of ontology rules using an RDF data store. We discuss various experiments based on spatial inference calculation. We discuss these results and present some solutions to improve the complexity of calculating spatial ontological inferences.**

***Keywords*—Spatial data model, Semantic data model, Trajectory data model, Spatial rules, Bussines rules, Ontology inference**

I. INTRODUCTION

Spatial database management systems are created to manage spatial data in terms of storing, computing relationships and querying. Several applications are based on spatial databases like geographic information systems (GIS), urban planning [5], route optimization [17] and traffic monitoring [14]. On the other hand, advances in information and communication technologies have encouraged collecting spatial, temporal and spatio-temporal data of moving objects [11]. Large databases need to be analyzed and modeled to meet the user's needs. However, to answer these queries we need to take into account the domain knowledge.

This paper deals with marine mammals tracking applications, namely seal trajectories. Trajectory data are captured by sensors included in a tag glued to the fur of the animal behind the head. The captured trajectories consist of spatial, temporal and spatio-temporal data. Trajectories data can also contain some meta-data. These sets of data are organized into sequences. Every sequence, mapped to a temporal interval, characterizes a defined state of the animal. In our application, we consider three main states of the seal: haulout, dive and cruise. Every state is related to the seal's activity. For example, the foraging activity occurs during dives. Although temporal aspects are important in such studies, we mainly focus here on the spatial dimension of these data.

We assume that our trajectory data are stored and managed in a spatial relational database. Then, we consider the query (Example 1) based on a schema (Code 1) of two spatial tables.

Example 1: Which dives are contained within a zone

```
Table Dive (idDive:integer, refSeal:string, maxDepth:real,
     shape:line(startPoint, endPoint));
Table Zone (idZone:integer, name:string, shape:polygon(
     points[]));
```

Code 1. Spatial schema

To answer the query (Example 1), we need a relational database language supporting spatial data. ISO/IEC 13249-3 SQL/MM [1] is the effort to standardize extensions for multi-media and application-specific packages in SQL. The standard is grouped into several parts. The part 3 [2] is the international spatial standard that defines how to store, retrieve and process spatial data using SQL. It defines how spatial data are represented, and the functions available to convert, compare, and process spatial data in various ways. Code 2 gives the SQL/MM expression of the query (Example 1).

```
SELECT D.idDive, D.refSeal
FROM  Dive D, Zone Z
WHERE Z.shape.ST_Contains(D.shape) AND Z.idZone = 5;
```

Code 2. The SQL/MM query of Example 1

The SQL/MM expression (Code 2) is based on a relational model of the trajectory data. This model represents the domain by a set of attributes and their values. Therefore, this model cannot take into account the domain knowledge as given by experts. We describe here for instance the query (Example 2).

Example 2: In which zones is the seal foraging

Even if the SQL/MM language provides spatial operations to solve the query (Example 1), it is not designed to resolve the query (Example 2). Indeed, the later query combines spatial data (`zone`), spatial operation (`contains`) and the semantic

domain knowledge related to the seal's activity (`foraging`). We notice that there is a semantic gap between the considered relational model of the trajectory data and the business process related to the domain knowledge as shown by the query (Example 2). Therefore, this paper addresses three main issues:

1) Trajectory domain model: The relational data model, in our trajectories data, is not suitable. Indeed, if for example we are interested to consider a generalization like (`Dive is a kind of Sequence`), all that the relational model can supply as a natural mechanism to express this constraint is a foreign key which concerns only the data and not the structuring links, as the generalization. In our work, an effective way to take into account the domain knowledge can be made through the user's needs. These needs are generally studied by the domain knowledge experts to formulate requirements or rules. As an example, the activity `foraging` is not a value or a set of values.
2) Spatial model: In the considered examples, the relations `dive` and `zone` can be assimilated to general spatial classes, respectively, `line` and `polygon`. Although, these are certainly not the only objects of the data model which can have spatial properties. Therefore, we believe that all spatial classes and properties must be considered regardless of the data model. The independent spatial model must be endowed with all the spatial reasoning.
3) Links between models: We based our approach on the definition of various separated models. Accordingly, we need to look at the problem of establishing links between these models. In the data engineering field, this problem is also known as data integration or mapping. Indeed, the study of this question is not recent and it arose from the need of reusing models.

This paper is organized as follows. Sections II illustrates the domain and semantic data models. Section III discusses OGC spatial data model implemented in different database systems and OGC spatial ontology model. Section IV details the implementation of the ontologies, the domain and spatial rules. Section V evaluates the spatial ontology inference over semantic trajectories. Section VI proposes enhancements over spatial ontology inference and experimental results evaluate the impact of them. Section VII summarizes some recent related work on spatial model and spatial semantics. Finally, Section VIII concludes this paper and presents some future prospects.

II. TRAJECTORY ONTOLOGY MODELING

A. Trajectory domain model

We consider trajectories of seals. The data are provided by LIENSs[1] laboratory in collaboration with SMRU[2]. These laboratories work on marine mammals' ecology. Trajectory data of seals between their haulout sites along the coasts of the English Channel or in the Celtic and Irish seas are captured using GNSS systems.

From the analysis of the captured data, we define a seal trajectory ontology that we connect to the trajectory domain ontology. The trajectory domain ontology is our model used in many moving object applications. Details of the modeling approach is discussed in [16]. Figure 1 shows an extract of the seal trajectory ontology, called `owlSealTrajectory`. Table I gives a dictionary of its concepts their relationships.

[1]http://lienss.univ-larochelle.fr - CNRS/University of La Rochelle
[2]SMRU: Sea Mammal Research Unit- http://www.smru.st-and.ac.uk

B. Seal trajectory model

In this work, we propose a Semantic Domain Ontology (Figure 2) based on activities organized as general ones linked to trajectory, and a hierarchy of basic activities linked to sequences of the trajectory domain ontology. The Seal Domain Ontology (Figure 2) is dealing with seal's activities. According to the domain expert, four activities (*resting, traveling, foraging* and *traveling-foraging*) are related to the three states of a seal. The seal trajectory ontology sequences are associated with these main activities.

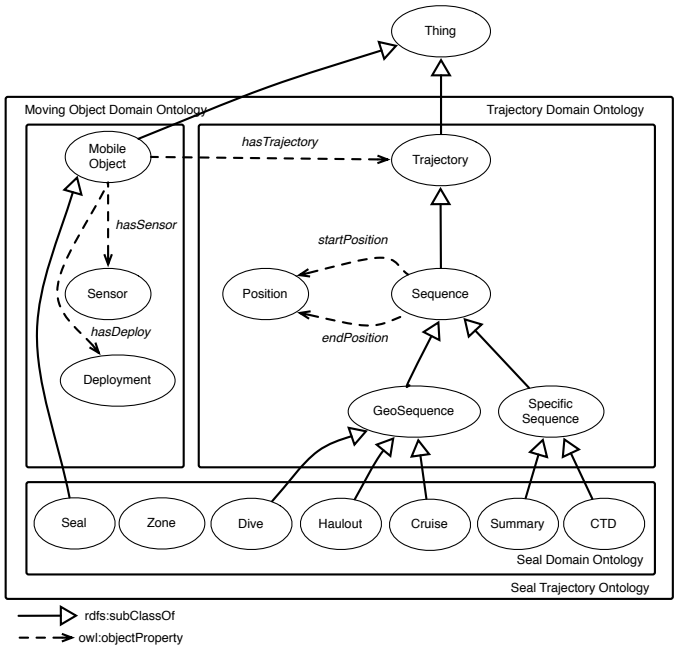


Figure 1. Overview of Seal Trajectory Ontology

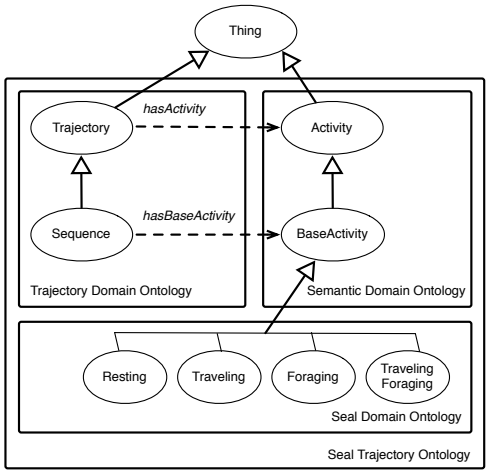


Figure 2. Overview of Seal Trajectory Ontology

Table I. SEAL TRAJECTORY ONTOLOGY DICTIONARY

Trajectory domain ontology	
Concept	**Description**
Trajectory	logical form to represent sets of sequences
Sequence	spatio-temporal interval representing a capture
GeoSequence	spatial part of sequence
Specific Sequence	metadata associated of a capture
startPosition, endPosition	object properties to represent the end and the beginning of a sequence
Seal domain ontology	
Concept	**Description**
haulout	a state of a seal when it is out of the water (on land) for at least 10 minutes
cruise	a state of a seal where it is in the water and shallower than 1.5 meter
dive	a state of a seal where it is in the water and deeper than 1.5 m for 8 seconds
Summary, CTD	metadata about deployment's conditions of the sensor, marine environment
dive_dur, sur_dur, max_depth	data properties: dive duration, surface duration and maximum depth of a dive, respectively
TAD	Time Allocation at Depth: data properties to define the shape of a seal's dive [5]

III. OGC SPATIAL DATA MODEL

We choose OGC to support spatial data, thanks for providing an OpenGIS simple feature specification for SQL [4]. This specification describes a standard set of SQL geometry types based on OpenGIS geometry model. Each spatial data is associated with a well-defined spatial reference system (SRID). SRID is a Spatial Reference IDentifier which supports coordinate system to uniquely identify any position on the earth. Latitudes and longitudes can be traced back to arbitrarily exact locations on the surface of the earth.

A. OGC geometry object model

The OGC geometry object model is based on extending the Geometry Model specified in the OpenGIS Abstract Specification. It is distributed computing platform neutral and uses OMT (Object Modeling Technique) notation. Figure 3 shows the object model for geometry. The base Geometry class has subclasses for Point, Curve, Surface and Geometry Collection. Each geometry object is associated with a Spatial Reference System (SRS) and has a Well-Known Text (WKT) presentation. Figure 3 shows aggregation lines between the leaf collection classes and their element classes. The OGC geometry object model defines relational operators on geometries. These are boolean methods that are used to test for the existence of a specified topological spatial relationship between two geometries. The specification is based on the Dimensionally Extended Nine-Intersection Model (DE-9IM) which describes the following kinds of spatial relationships: {`Touches`, `Crosses`, `Equals`, `Disjoint`, `Contains`, `Overlaps`, `Within`, `Intersects`}. The DE-9IM representation was developed by Clementini and others [7], [8] based on the seminal works of Egenhofer and others [9], [10].

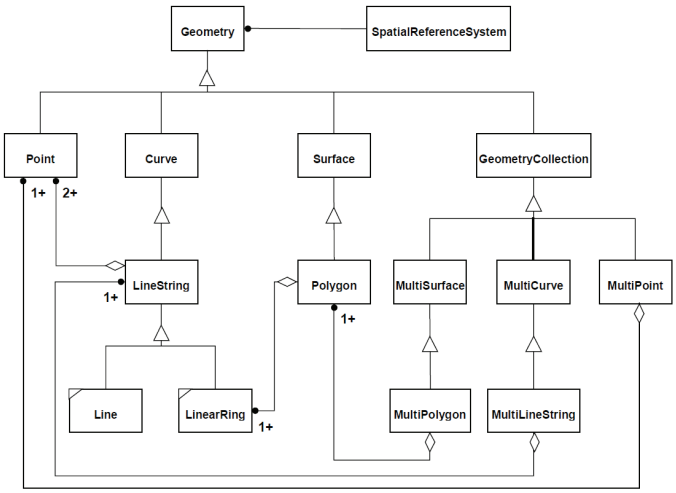


Figure 3. The OGC geometry object model hierarchy

B. OGC model in Oracle Spatial

Spatial supports the object-relational model for representing geometries. This model corresponds to an "SQL with Geometry Types" implementation of OpenGIS simple feature specification for SQL [4]. Spatial stores a geometry in Oracle native spatial data type for vector data, `SDO_GEOMETRY`.

The spatial relationship is based on geometry locations. The common spatial relationships are based on topology and distance. To determine spatial relationships between entities in the database, spatial has several secondary filter methods: `SDO_RELATE` operator evaluates topological criteria; `SDO_WITHIN_DISTANCE` operator determines if two spatial objects are within a specified distance of each other; `SDO_NN` operator identifies the nearest neighbors for a spatial object.

For example, the `SDO_RELATE` operator implements a nine intersection model for categorizing binary topological relationships between points, lines, and polygons. This yields to the set of spatial relationships:

```
SDO_covers, SDO_coveredby, SDO_contains, SDO_equal,
SDO_touch, SDO_inside, SDO_anyinteract, SDO_overlaps
```

Code 3. Topological relationships in Oracle Spatial

C. OGC spatial ontology

In our approach, we rewrite the OGC OMT class diagram (Figure 3) in UML class diagram. Then, we use model transformation techniques introduced by the Model Driven Engineering (MDE) community. For this, we choose an automatic transformation from UML class diagram into a formal ontology in OWL. We use transformer tool called *uml2owl Eclipse* [12]. This transformer, based on the metamodel *eCore Eclipse*, takes as input a UML class diagram and turns it into OWL-DL ontology. So, we transform the UML class diagram (Figure 3) to an OWL ontology, called `owlOGCSpatial`, Figure 4 presents an extract of it.

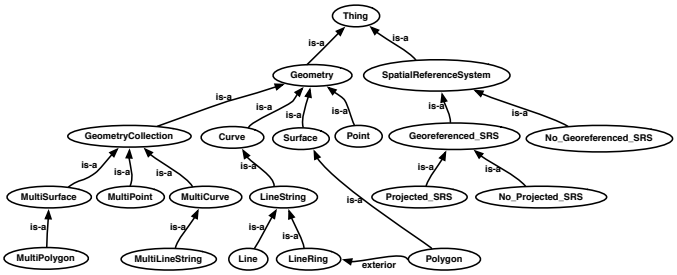


Figure 4. A view of owlOGCSpatial ontology

IV. Implementation of ontologies

A. Seal trajectory ontology rules

The seal trajectory ontology (Figure 1) is dealing with the seal's activities. Each seal activity has both a declarative part and an imperative part. The imperative parts of the activities are defined as rules in the ontology. A rule is an object that can be used by an inference process to query semantic data.

Oracle Semantic Technologies is a rule-based system where rules are based on IF-THEN patterns and new assertions are placed into working memory. Thus, the rule-based system is said to be a deduction system. In deduction systems, the convention is to refer to each IF pattern an antecedent and to each THEN pattern a consequent. SEM_APIS.CREATE_RULEBASE procedure defines user-defined rules in a rulebase. Our rulebase is called sealActivities_rb. The system automatically associates a view called MDSYS.SEMR_rulebase-name to insert, delete or modify rules in a rulebase. Code 4 gives foraging_rule definition based on domain expert's conditions. From line 4 to 10, we construct a subgraph and necessary variables needed by the IF part of the foraging_rule. Line 11 gives the THEN part of the rule. Line 12 defines the namespace of ontology.

```
1  EXECUTE SEM_APIS.CREATE_RULEBASE('sealActivities_rb')
2  INSERT INTO mdsys.semr_sealActivities_rb
3  VALUES( 'foraging_rule',
4  '(?diveObject rdf:type         s:Dive            )
5   (?diveObject s:max_depth      ?maxDepth         )
6   (?diveObject s:tad            ?diveTAD          )
7   (?diveObject s:dive_dur       ?diveDur          )
8   (?diveObject s:surf_dur       ?surfaceDur       )
9   (?diveObject s:seqHasActivity ?activityProberty )',
10 '(maxDepth > 3) and (diveTAD > 0.9) and                ( 
       surfaceDur/diveDur < 0.5)',
11 '(?activityProberty rdf:type  s:Foraging          )',
12 SEM_ALIASES(SEM_ALIAS('s','owlSealTrajectory#')));
```

Code 4. Implementation of the foraging rule

B. Spatial ontology rules

Open GIS specification considers two kinds of spatial relationships:

- Topological relationships based on the DE-9IM operators defined as methods on Geometry class: Equals, Within, Touches, Disjoint, Intersects, Crosses, Contains, Overlaps, Relate;
- Functions for Distance Relationships: Distance.

In this work, we consider topological relationships. Each relationship has a declarative part as an RDF, and an imperative part, formally, an associated rule as IF-THEN pattern. We create a rulebase named owlOGCSpatial_rb to hold spatial relationships rules. For example, the rule (Code 5) presents the imperative part of the spatial relationship Contains. Lines 4 to 10 in Code 5 represent the IF side of the rule. We construct a subgraph and necessary variables, namely, the two spatial objects sObj1 and sObj2, respectivity, their strings coordinates wktSObj1 and wktSObj2, and the srid which is the Spatial Reference System Identifier. The IF side of the rule evaluates the spatial relationship between the two spatial objects using a function called evalSpatialRelationship. This function builds a bridge between the ontology spatial rules and spatial operators in Oracle DBMS. Line 11 in Code 5 is the consequent or the THEN part of the rule.

```
1  EXECUTE SEM_APIS.CREATE_RULEBASE('owlOGCSpatial_rb');
2  INSERT INTO mdsys.semr_owlOGCSpatial_rb
3  VALUES('Contains_rule',
4  '(?sObj1    rdf:type    os:Geometry)
5   (?sObj2    rdf:type    os:Geometry          )
6   (?sObj1    os:srid     ?srid                )
7   (?sObj2    os:srid     ?srid                )
8   (?sObj1    os:wkt      ?wktSObj1            )
9   (?sObj2    os:wkt      ?wktSObj2            )',
10 '(evalSpatialRelationship(sObj1, wktSObj1, sObj2, wktSObj2
       , srid, ''Contains'' ) = 1)',
11 '(?sObj1    os:Contains   ?sObj2)',
12 SEM_ALIASES(SEM_ALIAS('os','owlOGCSpatial#')));
```

Code 5. Implementation of the Contains_rule

V. Spatial ontology inference on semantic trajectories

A. Ontology inference

Inferencing is the ability to make logical deductions based on rules defined in the ontology. Inferencing involves the use of rules, either supplied by the reasoner or defined by the user. At data level, inference is a process of discovering new relationships, in our case, new triples. Inferencing, or computing entailment, is a major contribution of semantic technologies that differentiates them from other technologies. In Oracle Semantic Technologies, inference process is based on entailments. We distinguish two entailments regimes [18]:

1) Standard entailment: there are several standard entailment regimes: semantics of RDF, RDFS and OWL. Support for RDF and RDFS is simplified by the availability of axioms and rules that represent their semantics. Support for major subsets of OWL-Lite and OWL-DL vocabularies have been provided. In this work, we use the subset OWLPRIME [22];
2) Custom entailment: since the standard vocabularies cannot handle all varieties of semantic application data, it becomes important to provide support for entailment based on arbitrary user-defined rules. In this work, we defined a rulebase for trajectory semantics sealActivities_rb and a rulebase for spatial relationships owlOGCSpatial_rb.

In Oracle Semantic Technologies, an entailment contains precomputed data inferred from applying a specified set of rulebases to a specified set of semantic models. Code 6 creates an entailment using seal trajectory and spatial models.

Other options are also required like number of rounds that the inference engine should run. In case of applying user-defined rules USER_RULES=T, the number of rounds should be assigned as default to REACH_CLOSURE.

```
1 SEM_APIS.CREATE_ENTAILMENT('owlSealTrajectory_idx',
2 SEM_MODELS('owlSealTrajectory','owlOGCSpatial'),
3 SEM_RULEBASES('OWLPrime','sealActivities_rb','
      owlOGCSpatial_rb'),
4 SEM_APIS.REACH_CLOSURE, NULL, 'USER_RULES=T');
```

Code 6. Entailment over the models and rullbases

B. Spatial ontology inference

The spatial ontology inference is the process of applying spatial ontology rules to compute topological relationships between spatial objects. Query 2, where we are looking for zones where the seal is foraging, combines the seal trajectory semantic Foraging and the spatial relationship Contains. To resolve it, the system needs an entailment over seal trajectory and spatial rules. The system must know the spatial relationships between zones and dives, considered as spatial polygons and lines, respectively. Figure 5 illustrates the computation algorithm of the inference over spatial objects. For every two spatial objects, the inference procedure calls spatial rules. The function evalSpatialRelationship calls the corresponding Oracle spatial operator for the current running spatial rule. The result of this function is returned to the spatial rule for checking the relationship between the two considered spatial objects. Computing a new relationship generates and saves a new inference triple.

VI. Enhance Spatial Inference

A. Restrictions and constraints refinement

The spatial ontology rules are computed redundantly to calculate the spatial relationships between geometries during the inference mechanism. To enhance the inference process, the user can define, for example, domain constraints limit the computation in a useful way for their work's objective. These limitations can be directional considering objects in the same direction or can be distance constraints considering a specific distance between objects or restrictions related to the type of the considered objects.

In our case, we define a refinement called *area of interest*. This refinement limits the computation of the inference to the objects located in a specified area. The area of interest refinement is given by Algorithm 1, considering two geosequences (S_a, S_r) and a given *area*. This algorithm gives the spatial relationship between the two considered geosequences S_a, S_r as output. This algorithm checks if these two geosequences belong to the interested area to compute the spatial relationship between them. If they do not belong to this area, the algorithm goes for another spatial candidate.

B. Passes refinement

We mention the REACH_CLOSURE problem in the case of applying user-defined rules. To control the number of the cycles done by the engine of Oracle during computing the inference, we define a refinement called *Passes refinement*. The passes refinement is illustrated by Algorithm 2 to effect the

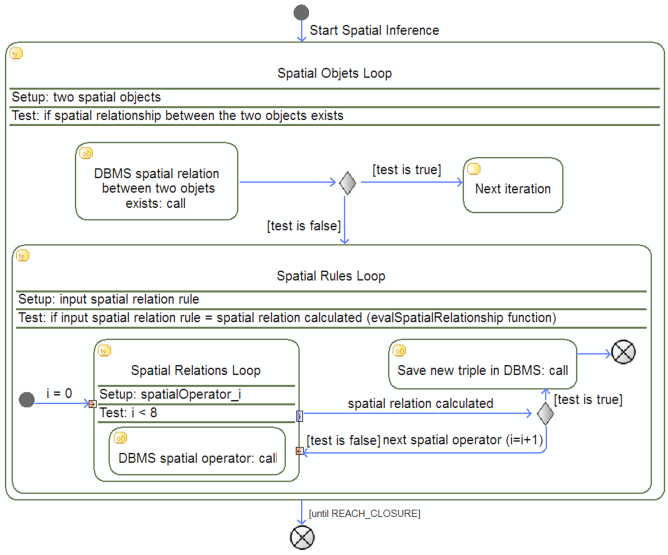


Figure 5. Activity diagram for spatial inference process

input : Two geosequences: a referent S_r and an argument S_a
input : An interested area *area*
output: Spatial relationship between S_r and S_a
initialization;
if $(S_a, S_r) \in$ *area* **then**
  calculate spatial relationship between S_r and S_a;
end
go the next geosequence S_{a+1};

Algorithm 1: Area of interest refinement algorithm

cycles of the engine. This algorithm takes the two considered geosequences (S_a, S_r) as input and provides the spatial relationship (*Res*) between (S_a, S_r) as output. This algorithm checks the computation of the inference between these geosequences if it exists. In the case of the inference passes for the first time, the inference process will be computed normally and its results will be given as output for this algorithm. In the other case where the inference is performed once before, the algorithm reads the saved result from the database and assigns it as a result to this pass. The function evalSpatialRules considers this refinement to optimize the passes of the engine during the computation of the inference.

input : Two geosequences: a referent S_r and an argument S_a
output: Spatial rule between S_r and S_a in *Res*
initialization;
if $(INFERENCE\ (S_a, S_r)) \in$ *database* **then**
  Res := result of the spatial rule from the database;
else
  Res := calculate inference between S_r and S_a;
  Save *Res* in the database;
end
go the next geosequence S_{a+1};

Algorithm 2: Passes refinement algorithm

C. Experimental results

In this section we evaluate the two spatial refinements we introduced in this work. In this evaluation, we consider sets of real seal trajectory data. The inference uses the eight spatial rules and the trajectory domain foraging rule, while the evaluation curves is given by the number of dives.

Firstly, we evaluate the area of interest refinement. Related to the seal trajectory domain and to our domain knowledge, we limit the area of interest restraint to 500 meters. We pass this candidate to Algorithm 1. The experimental results of this proposed refinement are shown in Figure 6. The results show its impact by the three following experiments:

1) *Spatial ontology rule calls - constraints refinement* presents the executions of the spatial ontology rules using the constraints refinement;
2) *Spatial ontology rule calls not executed* gives the reduced executions of the rules after the refinement;
3) *DMBMS spatial operator calls - constraints refinement* provides Oracle spatial operator calls during the inference process with the refinement.

We observe a decrease in both of the spatial ontology rules computation and DBMS spatial operator calls. For example, considering 250 dives, in the normal case of inference, the executions of the spatial ontology rules is 1000 000 and DBMS spatial operator calls is 125 000. However in the refinement case Figure 6, the executions of the rules is 130 000 and DBMS operator calls is 16 000.

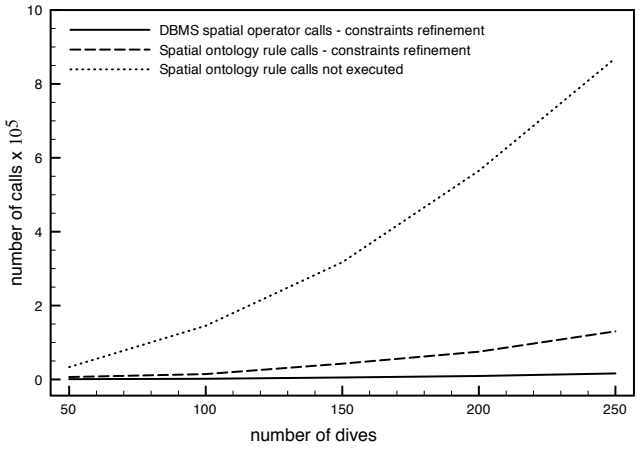


Figure 6. Enhancement of the spatial ontology inference with constraints refinement

Secondly, we evaluate the proposed passes refinement. The experimental results are shown in Figure. 7. The impact results are shown by the three following experiments:

1) *Spatial ontology rule calls* presents the spatial ontology rule calls during the inference process;
2) *Spatial ontology rule calls - passes refinement* displays the spatial ontology rule calls with the passes refinement;
3) *Spatial ontology rule calls - passes and constraints refinement* provides the spatial ontology rule calls with both refinements.

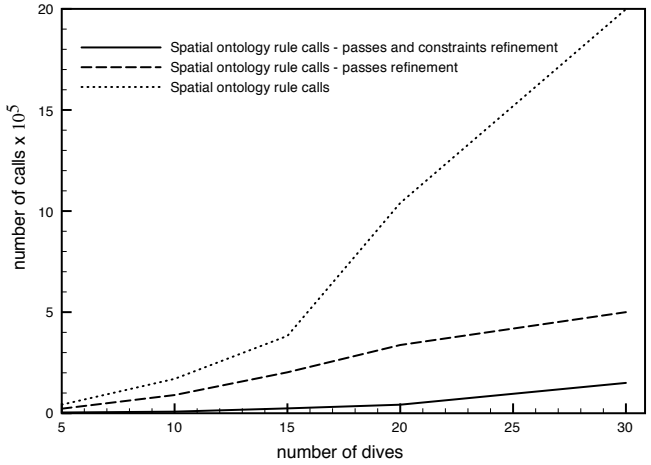


Figure 7. Evaluation of the spatial ontology inference over the proposed refinement

We observe a decrease in the spatial ontology rule calls with both refinements together. For example, considering 300 dives, in the normal case of inference, the spatial ontology rule calls is 2 000 000. However in the passes refinement case, the spatial ontology rule calls is 500 000. Finally in the both refinements case, the spatial ontology rule calls is 150 000. The final results are therefore considered as good impacts over the complexity and the size of the inference process.

VII. Related work

Several studies work on spatial domain, however so far there is no standard spatial ontology model. In 2007, the Geospatial Incubator Group (GeoXG), a W3C working group, tried to provide an overview of geospatial foundation ontology to represent geospatial concepts [15]. Moreover, GeoSPARQL ontology [13] represents and queries geospatial data on the semantic web. GeoSPARQL is based on OGC simple features model [4], with some adaptations for RDF. GeoSPARQL is a common query language for the Geospatial semantic web that can handle and index linked spatio-temporal data.

Enriching trajectory application domain with spatial model leads to manage semantics on trajectories. A conceptual view on trajectories is proposed by Spaccapietra et al. [19] in which the trajectories are a set of stops and moves. Stops are the important places of the trajectory where the object has stayed for a minimal amount of time. Vandecasteele et al. [20] adopted the trajectory data model proposed in [19] to detect abnormal ship behaviour by an enhanced spatial reasoning ontology. They integrated the spatial dimension into their ontology, and defined with their experts domain rules in Semantic Web Rule Language (SWRL). Nevertheless, the integration of the spatial dimension cannot yet be fully implemented in the ontology due to the lack of appropriate structures.

Moreover, by Battle et al. [3], a Geo-Ontology is an ontology design patterns for semantic trajectories. The authors used their semantic trajectory pattern to annotate two kinds of databases: trajectories generated by human travelers and by animals. This work lacks semantics and mainly do not support

inference over domain rules to enhance the semantic trajectories. The computational time taken by the inference mechanism including OWL-Time ontology is addressed by [21]. Based on a space-time ontology and events approach, Boulmakoul et al. [6] proposed a generic meta-model for trajectories to allow independent applications processing trajectories data benefit from a high level of interoperability, information sharing. Their approach is inspired by ontologies, however the proposed resulting system is pure database approach and a pure SQL-based approach not on semantic queries. Related to all those limitations, we design and implement an ontological trajectory framework integrated with the spatial dimension. The computation of domain and spatial rules as user-defined rules in our framework are the scope of this paper. We also propose some enhancements for the inference mechanism.

VIII. Conclusion and future work

In this work, we propose a modeling approach based on ontologies applied to the problem of thematic and spatial reasoning over trajectories. Our approach considers three separated ontology models: a general trajectory domain model, a domain knowledge or semantic model and a spatial domain model. We discuss the trajectory domain ontology and the semantic domain ontology. The considered semantic trajectory domain ontology connects the two previous models while considering moving object domain ontology. The spatial ontology model is based on the OGC standard. Therefore, we detail the specification of OGC Consortium and its different implementation in DBMS. To implement imperative part of the ontologies, we consider the framework of Oracle Semantic Data Store. To define the thematic and spatial reasoning, we implement rules related to the considered models. Thematic rules are based on domain trajectory activities and the spatial rules are based on spatial relationships. We compute the spatial ontology inference over semantic trajectories. Spatial rules directly influence the ontological inference process. This inference can be enhanced, so we address its main problems. For this reason, we propose some domain constraints and an inference refinement to enhance the spatial ontology inference. We evaluate our proposal on real trajectory data. The experimental results show the positive impact of the proposed approach. Finally, the objective of this paper is to extract in further details spatial characteristics revealed by and associated with moving object trajectory domain. So far, we used some domain application constraints over the ontological rules to effect positively the computation of the inference mechanism. For the future work, we would like to use a two-tier inference filters. In other words, two distinct operations are performed to enhance the inference: primary and secondary filter operations. The primary filter applies all the domain constraints over the captured data. In this paper, we consider a few of the domain interests, however for the future work, we will try to collect all the possible refinements, for example, analyzing data, classification or indexing. Then the primary filter permits fast selection of the filtered data or the analyzed data to pass along to the secondary filter. The latter computes the inference mechanism and yields the final knowledge data.

References

[1] ISO/IEC 9075-2:1999. In *Information Technology Database Languages SQL Part 2: Foundation (SQL/Foundation)*, 1999.

[2] ISO/IEC 13249-3:2002 FDIS. In *Information technology Database languages SQL Multimedia and Application Packages Part 3: Spatial*, 2002.

[3] R. Battle and D. Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012.

[4] D. Beddoe, P. Cotton, R. Uleman, S. Johnson, and J. R. Herring. OpenGIS Simple Features Specification For SQL. Technical report, OGC, 1999.

[5] E. Boeker. *Environmental Science; Physical Principles and Applications*. New York: Wiley, 2001. Ref. GE80.H69 1993.

[6] A. Boulmakoul, L. Karim, and A. Lbath. Moving object trajectories meta-model and spatio-temporal queries. In *International Journal of Database Management Systems (IJDMS)*, volume 4, pages 35–54, 2012.

[7] E. Clementini, P. D. Felice, and P. v. Oosterom. A small set of formal topological relationships suitable for end-user interaction. In *Proceedings of the Third International Symposium on Advances in Spatial Databases*, SSD '93, pages 277–295, London, UK, UK, 1993. Springer-Verlag.

[8] E. Clementini, J. Sharma, and M. J. Egenhofer. Modelling topological spatial relations: Strategies for query processing. *Computers & Graphics*, pages 815–822, 1994.

[9] M. J. Egenhofer and R. Franzosa. Point-set topological spatial relations. Number 5(2), pages 161–174, 1991.

[10] M. J. Egenhofer and J. Herring. A mathematical framework for the definition of topological relationships. pages 803–813, 1990.

[11] R. Güting and M. Schneider. *Moving Objects Databases*. Morgan Kaufmann, 2005.

[12] G. Hillairet, F. Bertrand, and J. Y. Lafaye. MDE for publishing data on the semantic web. In *international workshop on Transformation and Weaving Ontologies and Model Driven Engineering (TWOMDE) at MODELS'08*, 2008.

[13] Y. Hu, K. Janowicz, D. Carral, S. Scheider, W. Kuhn, G. Berg-Cross, P. Hitzler, M. Dean, and D. Kolas. A geo-ontology design pattern for semantic trajectories. In *Spatial Information Theory*, volume 8116 of *Lecture Notes in Computer Science*, pages 438–456. Springer International Publishing, 2013.

[14] F. Korn, S. Muthukrishnan, and Y. Zhu. Checks and balances: monitoring data quality problems in network traffic databases. In *Proceedings of the 29th international conference on Very large data bases - Volume 29*, VLDB '03, pages 536–547. VLDB Endowment, 2003.

[15] J. Lieberman, R. Singh, and C. Goad. W3C Geospatial ontologies - W3C incubator group, 2007.

[16] W. Mefteh. *Approche ontologique pour la modelisation et le raisonnement sur les trajectoires. Prise en compte des aspects thematiques, temporels et saptiaux*. PhD thesis, La Rochelle university, 2013.

[17] S. Sangheon, Pack Xuemin sherman, M. Senior, W. M. Jon, F. Life, and P. Jianping. Adaptive route optimization in hierarchical mobile IPv6 networks.

[18] D. Souripriya and S. Jagannathan. Database technologies for rdf. In *Reasoning Web. Semantic Technologies for Information Systems, 5th International Summer School 2009*, volume 5689 of *Lecture Notes in Computer Science*, pages 205–221. Springer, 2009.

[19] S. Spaccapietra, C. Parent, M. Damiani, J. Demacedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data and Knowledge Engineering*, pages 126–146, 2008.

[20] A. Vandecasteele and A. Napoli. An enhanced spatial reasoning ontology for maritime anomaly detection. In *International Conference on System Of Systems Engineering - IEEE SOSE*, GIS '06, pages 247–252, 2012.

[21] R. Wannous, J. Malki, A. Bouju, and C. Vincent. Time integration in semantic trajectories using an ontological modelling approach. In *New Trends in Databases and Information Systems*, volume 185 of *Advances in Intelligent Systems and Computing*, pages 187–198. Springer Berlin Heidelberg, 2013.

[22] Z. Wu, G. Eadon, S. Das, E. I. Chong, V. Kolovski, M. Annamalai, and J. Srinivasan. Implementing an inference engine for rdfs/owl constructs and user-defined rules in oracle. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 1239–1248, Washington, DC, USA, 2008. IEEE Computer Society.

Large Scale Desalination: A Comparative Cost Affective Economic Analyses Of Nuclear, Gas and Solar Powered Plants

Mohammed H. S. Al Ashry
Shaqra University
The Community College

Abstract: *The main objective, here, is to explore the economic viability of the solar powered desalination method through a cost and benefit comparative and contrast study. Using the initial construction expenditure, the annual maintenance cost, and energy consumed or produced a variance ratio test of the random walk hypothesis will be implemented to determine their relative financial efficiency. This paper will also utilize the first order autoregressive multivariate estimation model to analyze the methods and identify the most productive process with most financial promise for future investment. The total deviations of the estimated variables from the actual are accounted for by the variations of the variances of the estimates from the actual. The higher the percentage of the unexplained deviation the higher the risk involved. The portfolio variance will be utilized to measure the investment risk in the three desalination industries.*

Index:
System's lifecycle: the time after which the system become a liability
Heteroscedasticity: different sampling variables
Autoregressive: descriptive estimates of random variables
Asymptotic: a line infinitely approaching another curve-linear lines

Glossary of Terms:
$\check{R}_{i,t}$ = estimated variable
θ_i = estimation constant
π_i = variables' estimation coefficient
$\varepsilon_{i,t}$ = random variable above or below the average variable's estimate
r = correlational coefficient
σi = variance / covariance
V(Ři) = portfolio variance variable
$\mathbf{R}^2$ = coefficient of determination
ŋ, $_g$ and đ = sampling percentage ratios

1. Introduction:

Water is probably the most important commodity affecting the lively-hoods of the majority of the populations on earth. In due course, water will be at the center-stage of a worldwide crisis that may consume millions of lives. It is imperative to find the least expensive and most productive approach to mass produce water in order to satisfy the future needs of earth's inhabitants. Not much literature is written on direct solar energy for the purpose of producing drinking water, not on a large scale anyway.

The nuclear, gas and solar energy schemes differ in the technique to produce energy [1], [2], however, there are few viable options to produce drinking water. In this paper only the Multi-Stage-Flash (MSF) system is considered for the desalination of seawater [3]. MSF uses a process in which seawater is heated, evaporated then condensed to produce drinking water. For the nuclear and gas plants the water, on average, is heated to boiling temperature, 100 - 105 C0, to increase the percentage of the evaporated water, which in turn increases the percentage of condensed water. However, the downside, in this case, is that increasing the heated water temperature, on the long run, lowers the efficiency of MSF system, and reduces the span of the system's lifecycle. This increases cost and lowers its investment potential. The direct solar heat, on the other hand, heats the seawater in long parallel ducts, figure-1, bottom of the table, to a reasonable temperature subject to the location and seasonal temperature variations. In Saudi Arabia the water collectors' temperature averages about 80 C0 for the solar-heated seawater during the days of the summer months [7]. The average temperature difference between the seawater and the seaside ground surface fluctuates for nights and days, winter to summer periods, 30 to 50 C0, respectively [7]. This increases the evaporation and condensation during the summer period, however, the temperature interval separating the lows and highs during the rest of the year is enough for the production of large amounts of drinking water [3], [7].

In Saudi Arabia a large number of gas-powered seawater desalination plants are either, operating, being built, or planned. Due to the scarcity of water in the Arabian peninsula Saudi Arabia may rely on seawater desalination for a long time to come. However, it is worth mentioning that on the long run, although not economically proven feasible, some believe that nuclear energy maybe the most reliable for, simultaneously, producing and desalinating electricity and sea water, respectively [4], [8].

The data can be divided into informational data such as the cost of energy, operation and maintenance (O&M) per million joules (MJ), and the cost of construction ($)/MJ. This information is used to calculate the total cost per million joules (TC/MJ) of energy [1]. Maintenance includes energy, labor, parts and other indirect costs. Construction includes, for nuclear power, nuclear reactors, turbines, heaters, all

other required plant's concrete chamber(s), compressors, gauges and monitors; for gas power, all of the above, excluding the nuclear reactor, and adding the gas-based electric turbines (6).

1.1 Defining the parameters: The data for the nuclear and gas energy based plants are extracted from US energy production costs between 1995 and 2011. The previous three years were the result of a regressive inflation based extrapolation [1] [2]. However, the solar powered plant's data is entirely an inflation-based extrapolated estimates [7]. (TC/MJ) is utilized for two purposes:

a. To calculate the profit under the assumption that the initial cost is financed. In this case the cost of the first year is multiplied by 2000 to obtain the present worth (PW) of the financing money for the establishment of a 2GJ energy-plant, for the production of drinking water. The annual payment (AP) and the future worth (FW) of the borrowed money is calculated at the bottom of tables I, II & III for each of the industrial sectors, however only (table-I) will be discussed since all tables follow the same mathematical procedure, except for the solar energy where the data is extracted entirely, as mentioned above, from inflation based data, and is the main subject of this paper. The total cost of a two Giga Joules of an energy plant will be used as the present worth of the project to estimate and gauge their future value over twenty years. The first year, 1992, will be used to calculate the present worth-value to finance the project. The financing data will be used purely for comparative purposes with the actual data and its estimates to emphasize the viability of the marketing future of the desalination industry.

b. The actual total of the varying inflation-based cost [13], will be calculated by tallying the cost of the first year, 1992, plus the aggregate portions greater than the 1992-magnitude, for the duration of the time series to 2011. With the exception of the nuclear sector, both the gas-based and the direct sun-radiation energy-based sectors fluctuated above and below the 1992 magnitude, due to many factors which will be discussed later. The future value of the actual total cost is displayed in row 25, under the "Actual total cost/MJ", in table I, T1, T2 and T3; for the three sectors, respectively, where Ti= [(Ri*2000)+ ∑(Rij - Ri) for i= 1 and j= i ..n; where i….n stands for the magnitudes of 1992 – 2011. These values for the three sectors are also solely as reference and for comparative purposes. The actual market rates' time series values for the three sectors, 'revenue data', (R1, R2 and R3), will be estimated using the first order autoregressive model (FOA) to estimate Ř1, Ř2 and Ř3 [12]. The actual and estimated values will be utilized to run a variance ratio test of the random walk hypothesis of both variances, table II, VR(q_a) and VR(q_e). This is employed to assess the viability of the energy producing schemes including the solar seawater desalination. The data in turn can be evaluated using the variances and data fluctuations of the three energy schemes. The objective is to forecast the viability of the solar energy for seawater desalination. The notion that annual changes in the cost of energy-projects including desalination are equivalent to changes in stocks' earning yields is adopted to facilitate the process. The stocks of the involved manufacturing businesses are rising due to people's growing need for drinking water. The increasing cost of seawater-desalination projects is due to both, the rising cost of the energy and the emergence of numerous and diverse advances in desalination technology that have not been extensively tested; despite its competitive market. Energy projects for water desalination are investments known for its high return, however the risk associated with such projects adds to the uncertainty of the seawater desalination ventures in general. Solar seawater desalination eliminates this uncertainty [7]. The cost-effectiveness and efficiency of the desalination of seawater through direct solar energy is an asset that someday will provide the world with most of its drinking water.

2. Application analysis:

The schemes employed are intended to test for the most appropriate energy-based-process for global mass water-production. In this paper, the total per energy cost of the construction, operation and maintenance is subject to changing inflation, depreciation, and rate changes in the cost of material due to constantly changing technology. However, the cost is actually an investment, part of which is earnings to investors; which maybe a cost to the project's owners hoping to make a profit on the long run. There are three parties involved here, the manufacturers, for whom the project is revenue, the banks or investors providing the loan, and the owners operating or leasing the finished project for profit. The first year's Cost/MJ of energy is used as the base value to calculate the present value of the investment. The financing uses this value to calculate the annual payment and future value of the investment after twenty years with a 3% interest rate, purely for economic comparison purposes. The calculated future values (FW), as shown in table I, for the solar case, with annual principle payment AP1, FW ranges from 7 to 9% of the actual annually changing per energy cost T1. This indicates a successful endeavor despite the low interest rate. The solar desalination actual cost of the energy is lowest in comparison to the nuclear and gas powered methods. However, that is not necessarily enough to appreciate this scheme. The cost of such technique may be higher in other countries. Nations in the northern and southern hemisphere may not have enough solar radiation to generate enough heat, not to mention the sea-side ground elevation relative to seawater level [7]. Nations near the equator especially in desert areas may not have the appropriate geography to employ such approach, and altering the landscape maybe too costly.

3. Technical analysis:

3.1 The methodology:

This paper uses the total annual Cost/MJ for nuclear, gas and solar seawater desalination industries, (table I, for the solar industry), as industrial investments' annual revenues,

Table I: Energy Cost Comparison

Solar			Estimated cost/MJ =AP_3+ Escrow	the weight đ	E(R_3)	σ_{a3}^2	đσ^2	Θ_{ij}	$\prod_{ij}$	Eq-2 $\check{R}_3(e)$	σ_{e3}^2	$V_j(R_i)$ weighteσ_{e3}^2	Eq-7 đVe(R3)	Port-folio-3	total actual project cost
	Actual Total cost/MJ =R3	average $\check{R}_3$													
	0.0005312	0.00057	0.0959	1.21414E-05	0.52	4E-09	5E-14	6E-04	1.97E-11	1205	5.5E+04	0.66966	26.786248	970188795	0
	0.0005412	0.00057	0.0959	1.21E-05	0.51	4E-09	5E-14	6E-04	1.96E-11	1236	5.5E+04	0.66702	26.680663	972820809	1E-05
	0.0005413	0.00057	0.0959	1.18E-05	0.494	4E-09	5E-14	6E-04	1.96E-11	1266	5.5E+04	0.65256	26.1023	975337720	1.01E-05
	0.00054164	0.00057	0.0959	1.15E-05	0.482	4E-09	5E-14	6E-04	1.95E-11	1303	5.5E+04	0.63669	25.467502	978147755	1.04E-05
	0.000552	0.00057	0.0959	1.15E-05	0.442	4E-09	5E-14	6E-04	1.94E-11	1342	5.5E+04	0.6337	25.347944	984075496	2.08E-05
	0.000552	0.00057	0.0959	1.12E-05	0.506	4E-09	5E-14	6E-04	1.95E-11	1374	5.5E+04	0.61525	24.609866	978401620	2.08E-05
	0.000552	0.00057	0.0959	1.12E-05	0.311	4E-09	5E-14	6E-04	1.91E-11	1396	5.5E+04	0.62026	24.810253	1001670727	2.08E-05
	0.0005518	0.00057	0.0959	1.1E-05	0.335	4E-09	4E-14	6E-04	1.91E-11	1427	5.5E+04	0.6051	24.203765	999452953	2.06E-05
	0.00054164	0.00057	0.0959	1.04E-05	0.333	4E-09	4E-14	6E-04	1.91E-11	1477	5.5E+04	0.57491	22.99629	1001247052	1.04E-05
	0.0005315	0.00057	0.0959	9.88E-06	0.394	4E-09	4E-14	6E-04	1.92E-11	1520	5.5E+04	0.54483	21.793055	995198007	3E-07
	0.0005216	0.00057	0.0959	9.61E-06	0.342	4E-09	4E-14	6E-04	1.90E-11	1544	5.5E+04	0.52981	21.192269	1002004567	-9.6E-06
	0.0005316	0.00057	0.0959	9.62E-06	0.301	4E-09	4E-14	6E-04	1.89E-11	1580	5.5E+04	0.53086	21.234306	1007945410	4E-07
	0.0005317	0.00057	0.0959	9.46E-06	0.238	4E-09	4E-14	6E-04	1.88E-11	1624	5.5E+04	0.52162	20.864732	1017427455	5E-07
	0.00054175	0.00057	0.0959	9.24E-06	0.302	4E-09	4E-14	6E-04	1.89E-11	1681	5.5E+04	0.50977	20.390567	1010168091	1.06E-05
	0.000552	0.00057	0.0959	9.07E-06	0.349	4E-09	4E-14	6E-04	1.90E-11	1737	5.5E+04	0.50044	20.017566	1005922605	2.08E-05
	0.0005725	0.00057	0.0959	9.12E-06	0.384	4E-09	4E-14	6E-04	1.90E-11	1788	5.5E+04	0.50291	20.116295	1003328712	4.13E-05
	0.0006337	0.00057	0.0959	9.67E-06	0.43	4E-09	4E-14	6E-04	1.91E-11	1860	5.5E+04	0.53358	21.342815	1000279624	0.000103
	0.0006706	0.00057	0.0959	1.02E-05	0.488	4E-09	4E-14	6E-04	1.92E-11	1853	5.5E+04	0.56339	22.535359	994761454	0.000139
	0.0007052	0.00057	0.0959	1.05E-05	0.539	4E-09	4E-14	6E-04	1.93E-11	1884	5.5E+04	0.58052	23.220461	991055957	0.000174
	0.0007453	0.00057	0.0959	1.08E-05	0.575	4E-09	4E-14	6E-04	1.93E-11	1943	5.5E+04	0.5941	23.763745	989500753	0.000214
	0.001	AP3	0.0714							1552			Risk	-1.99%	2.69886
	23	PW	1.0624												
		FW	1.9188							55155					
Std	6.36863E-05		0.0959							3.E+04					
StdAv	0.111318002		140.7%		$\Theta_{ij} = \check{R}_i - r\,(\sigma_i / \sigma_j)\,\check{R}_j$ for i = 1 & 3					e stands for estimated					
		actual variance	4.056E-09		$\prod_{ij} = r_{ij}\,(\sigma_i / \sigma_j)$ for j = 3 & 1					a stands for actual					
			0.083848							AP=annual payments =0.0714 -------				0.0714	
										PW = present worth =1.0624 --------				1.0624	
										FW = future worth =1.9188 --------				1.9188	
										Estimated cost/MJ =AP3 + Escrow= ------				2.69886	
										Ratio of actual to estimated revenue =T1/FW				140.7%	

Ri, with R1, R2 and R3, for the industries mentioned above, respectively. The data is utilized to assess the viability of solar seawater desalination as a continuous source of drinking water. The actual data is estimated using the first order autoregressive model equations-1-6, table I, [5][12]. Equation-7 is used to test for the stability and deviation of the explained and unexplained autoregressive model's variations between the actual and estimated variances table I. Then using equations-8-11, the variance ratio (VR) is calculated to test for the random walk hypothesis, table II. Finally, equations 8 and 13, the heteroscedasticity asymptotic variance (HAV) and 15, the consistency form, to test for the market - earnings consistency, in this case growth or a bubble; table II. Equation-16 tests the estimated variables and variances for investment risks for the three industrial sectors. The equations mentioned above are listed in the order they are employed as follows:

$$\check{R}_{i,t} = \theta i + \pi i\, R_{i,t-1} + \varepsilon_{i,t} \qquad i = 1..n \quad (1)$$

Where Ri,t = the values vector represented by the industrial return at time t. "θi, πi = constants calculated using ordinary least squares model $\varepsilon_{i,t}$ = a noise random variable in the industrial-return variables, averaging zero with a value of 1 for standard deviation in bivariate dependent regression, which maybe more instrumental in economics sensitivity studies; overall can be eliminated in earnings estimates."[5]. Equation-1, puts more emphasis on regressive estimates of the actual values with emphasis on the so called unexplained disturbance (noise) for multivariate regression. This unexplained disturbance, in reality is an error that can be explained by the variation in the deviation of the estimated variance from that of the actual over the regression period. In this case, it is simplified by employing the time series periods and degrees of freedom to factor into the co-variances instead of using expected return coefficients. The coefficients in this equation correlate the independent variables.

$$\check{R}_{i,t} = \theta i + \pi i\, R_{i,t} \qquad i = 3, 1 \quad (2)$$

Where $\theta i = \check{R}3 - r\,(\sigma_i / \sigma_j)\,\check{R}1$ (3)

(for solar desalination i = 3)

if i = 3, 1 then (σi / σj) = (σ3 / σ1), (σ1 / σ3), i≠ j, respectively.

$$\pi i = r\,(\sigma 3 / \sigma 1) \qquad \text{for } i = 3 \quad (4)$$

$$V(Ri) = \pi i2\, V(\check{R}i) + V(\varepsilon_{i,t}) =$$

$$(r\,(\sigma i / \sigma j))^2\, V(\check{R}i) + V(\varepsilon_{i,t}) \quad (5)$$

Table II: Variance Ratio

Factors		Total cost/MJ actual	Total cost/MJ estimated	Total cost/MJ	Total cost/MJ		heteroscedistic asymptotic	Z - test
						VR(q)a	variance (HAV)	
Time↓	Fuel	$X_t=R_1:R_3$	$X_t(e)=R_1:R_3$	$(X_{t-q})a$	$(X_{t-q})e$	V(Xt-k)/V(Xt)	*v*(k)a*	Z*(k)a
1992		3.9000E+01	1.1116E+03	4.1000E+01	1.3701E+08	0.966356446	3.84329E+18	-4.4E-21
1993		4.0000E+01	1.1424E+03	4.4484E+01	1.4422E+08			
1994		4.1000E+01	1.1733E+03	4.7812E+01	1.1972E+08			
1995		4.2170E+01	1.2095E+03	5.1161E+01	1.1741E+08	VR(q)e	*v*(k)e*	Z*(k)e
1996		4.3443E+01	1.2487E+03	5.6236E+01	1.3269E+08	0.998432606	2.15787E+17	-7.5E-20
1997		4.4484E+01	1.2811E+03	6.0000E+01	1.6298E+08			
1998		4.5184E+01	1.3021E+03	4.7504E+00	1.3699E+08			
1999		4.6196E+01	1.3335E+03	4.7504E+00	1.3699E+08			
2000		4.7812E+01	1.3834E+03	3.9004E+00	1.1248E+08			
2001		4.9205E+01	1.4267E+03	4.6008E+00	1.3268E+08			
2002		5.0000E+01	1.4511E+03	3.6507E+00	1.0528E+08			
2003		5.1161E+01	1.4868E+03	4.9007E+00	1.4133E+08			
2004		5.2570E+01	1.5301E+03	6.0008E+00	1.7306E+08			
2005		5.4414E+01	1.5874E+03	5.4120E-04	1.2355E+03			
2006		5.6236E+01	1.6439E+03	5.5200E-04	1.3419E+03			
2007		5.7886E+01	1.6950E+03	5.5180E-04	1.4269E+03			
2008		6.0204E+01	1.7668E+03	5.2160E-04	1.5444E+03			
2009		6.0000E+01	1.7606E+03	5.4175E-04	1.6807E+03			
2010		6.1000E+01	1.7917E+03	6.3370E-04	1.8596E+03			
2011		6.2891E+01	1.8503E+03	7.4530E-04	1.9426E+03			
1992		4.7504E+00	-1.6494E+04	1.7540E+01	8.7642E+12			
1993		4.7504E+00	-1.6494E+04					
1994		4.7505E+00	-1.6494E+04					
1995		4.7504E+00	-1.6494E+04					
1996		4.6005E+00	-1.6494E+04					
1997		5.0005E+00	-1.6493E+04					
1998		3.9004E+00	-1.6494E+04					
1999		4.1005E+00	-1.6494E+04					
2000		4.1508E+00	-1.6494E+04					
2001		4.6008E+00	-1.6494E+04					
2002		4.3005E+00	-1.6494E+04					
2003		4.0707E+00	-1.6494E+04					
2004		3.6507E+00	-1.6494E+04					
2005		4.2008E+00	-1.6494E+04					
2006		4.6007E+00	-1.6494E+04					
2007		4.9007E+00	-1.6493E+04					
2008		5.3008E+00	-1.6493E+04					
2009		5.6508E+00	-1.6493E+04					
2010		6.0008E+00	-1.6493E+04					
2011		6.3008E+00	-1.6493E+04					
1992		5.3120E-04	1.2046E+03					
1993		5.4120E-04	1.2355E+03					
1994		5.4130E-04	1.2664E+03					
1995		5.4164E-04	1.3025E+03					
1996		5.5200E-04	1.3419E+03					
1997		5.5200E-04	1.3740E+03					
1998		5.5200E-04	1.3956E+03					
1999		5.5180E-04	1.4269E+03					
2000		5.4164E-04	1.4768E+03					
2001		5.3150E-04	1.5198E+03					
2002		5.2160E-04	1.5444E+03					
2003		5.3160E-04	1.5802E+03					
2004		5.3170E-04	1.6238E+03					
2005		5.4175E-04	1.6807E+03					
2006		5.5200E-04	1.7370E+03					
2007		5.7250E-04	1.7880E+03					
2008		6.3370E-04	1.8596E+03					
2009		6.7060E-04	1.8533E+03					
2010		7.0520E-04	1.8842E+03					
2011		7.4530E-04	1.9426E+03					
average		1.8320E+01	-4.4943E+03					

Equation-5 can be employed for estimating variance of two variables however our case requires equation-7 due to the extreme fluctuations of our industrial variables. Equation-2, estimates the actual variables' for two variables. In this process, as mentioned in the previous paragraph, the noise' standard deviation is 1, and averaging zero which allows economists to exclude the disturbance, in this case, when estimating revenue. However, equation-1 is more appropriate where the unexplained deviation is formulated by manipulating equation-7. For example, equation-6 is an example of a partial correlation coefficient where r_{23} correlates the independent variables R_2 and R_3, a correlation coefficient for a third variable is utilized to compensate for the significant disparities in magnitude between the separate variables, So for i =2, j=3 $i \neq j$

$$\pi_i = r_{ij}(\sigma i/\sigma j) = r_{32}(\sigma_3/\sigma_2) \ \& \ r_{31}(\sigma_3/\sigma_1) \qquad (6)$$

or vise versa for the variance and correlation coefficients.

$$V(Ri)=\mathbf{R}^2V(\check{R}i)\}+V(\varepsilon_{i,t})= [V(\check{R}i)(k-1)]/[V(\check{R}i)(k-1)+(n-k)cov(\check{R}_i,\check{R}_a)]$$

$$V(\check{R}i)\}+V(\varepsilon_{i,t}) = [(k-1)\ (\sigma i2)]/ [(k-1)(\sigma_i^2)+(n-k)(r_{ai}\,\sigma i\,\sigma_a)]V(\check{R}i) + V(\varepsilon_{i,t}) \qquad (7)$$

In this equation we can see the coefficient of determination $\mathbf{R}^2$ which is a variance ratio used primarily to simplify explaining the variance deviation between the actuals and their estimates, and is instrumental in calculating the noise [10].

The following, equation-8 is the population variance ratio relating the variables on the basis of their degrees of freedom to accentuate the sampling population variability, where the fluctuation in equation-9 is examined in relationship to equation-10.

$$VR(k) = \sigma^2(k)/\sigma^2(1) \quad (8)$$

$$\sigma^2(k) = 1/k(nk-k+1)(1-(k/nk))^* nk\Sigma_{t=k}(X_t - X_{t-k} - k\ddot{x})^2 \quad (9)$$

$$\sigma^2(1) = 1/(nk-1)nk\Sigma_{t=1}(Xt - X_{t-1} - \ddot{x})^2 \quad (10)$$

Where $\ddot{x} = (1/nk)(x_{nk} - x_0)$ (11)

$$v(k) = 2(2k-1)(k-1)/3k(nk) \quad (12)$$

Equation-13 is similar to equation-8 with more emphasis on the heteroscedasticity of the variance variables; in this case, not only is the sampling of the population is important, but also the diversity of the variables.

$$v^*(k) = {}^{k-1}\Sigma_{m=1}[2(k-m)/k]^2 [nk\Sigma_{t=m+1}(X_t - X_{t-1} - \ddot{x})^2 (X_{t-m} - X_{t-m-1} - \ddot{x})^2]/ [nk\Sigma_{t=m+1}(Xt - Xt\text{-}1 - \ddot{x})^2]^2 \quad (13)$$

$$Z(k) = (VR(k)-1)/(v(k)^{-(1/2)}) \quad (14)$$

$$Z^*(k) = (VR(k)-1)/(v^*(k)^{-(1/2)}) \quad (15)$$

$$V(R) = \eta^2 V(R_1) + {}_g{}^2 V(R_2) + đ^2 V(R_3) + 2\eta_g r\sigma_1\sigma_2 + 2\eta đ r\sigma_1\sigma_3 + 2_g đ r\sigma_2\sigma_3 \quad (16)$$

Equations-14 and 15 relate the variance ratios for homoscedasticity and heteroscedasticity, respectively. Equation-16 is the portfolio selection theory formula [11]; it provides the industrial investment risk assessment employing the variables and variances' estimates.

3.2 The abstract analysis:

The solar-heat plant's maintenance's cost encompasses cleaning the reservoirs' walls and surfaces, the condensers and evaporators piping along with the solar fields metallic surfaces and concrete inner ducts. The cleaning is done through flush-rinsing the surfaces, ducts and pipes with forced water and using water suctioning to help clear the surfaces and linings of precipitants. The cost is estimated to be $1 per square meter plus $1 per square meter for coating the solar fields' metal surfaces with less than one millimeter of tar. The total surface area including the ducts and piping is estimated to be 10 Mm2. (million square meters). The total cost = 10*1 + 4(total tar surface-area)*1 = $14 M; the cost pricing employs the market's whole-sale maintenance pricing in Saudi Arabia during the late Nineties. The cost per square meter = 14/10 = $1.4 per square meter and the cost of the energy produced = 14ex6/ (2.96*4ex6) =1.18 per MW, where the energy produced ~= 2.96 MJ/m2 and the heat collectors area is 4 million square meters [7]. The inflation index between 1995 and 2013 has a low of 98 with a high of 148. Using an average approximation we get 2.8 as the annual (consumer price index) average inflation, with the year 1999 as the base index for the actual data on table I. Table(III): sums-up the results of the analyses of the three industries and their feasibility on the bases of the tests done above.

Table III: The analyses				
Variables		Solar	Gas	Nuclear
The random walk hypothesis	VR (q) = 1	The result of this test VR(q)=0.97 (actual) and =.99 (estimated) emphasizes that the null hypothesis VR (q) = 1 indicates that the selected index of business sectors follows a random walk. This means that the values of the sector are influenced by specific internal or external factors that can be manipulated by individuals or other sectors. This can be asserted by the fact that the main factors affecting the desalination process are not controlled by one or two entities. The main factor is the desalination technology itself, where the utilization of refined high resolution technologies such as the nanotechnology and material amalgamation led to the production of highly refined purification schemes and temperature controlled piping systems. Energy source utilization technology, for example is diverse where we have, in addition to the three mentioned above, coal, oil, wind and other forms of hydroelectric energy sources. In each of these power producing mechanisms the technology is advancing at a fast pace. Fuel saving, emission controls and efficiency are major factors influencing decision makers. Energy sources such as gas, nuclear, solar, etc. are another factor that can be influenced by politics, delivery and distribution and related shipping safety, international relationships, refining technology and more. Desalination technology is more independent than the energy or the energy utilizing technologies. Due to the versatility of the energy techniques employed for desalination and the diversity of the desalination mechanisms, it is that such index does follow a random walk hypothesis		
	VR (q) < 1	Although VR (q) ~= 1, it is possible the hypothesis is rejected returning VR (q) < 1 which indicates that the related sectors are negatively serially correlated and in a sense in the process to correct a mature market. This interpretation is also possible. Desalination and its related energy sources and technologies have been around for a long time, and the market has gone through many changes and fluctuations technologically and financially. Another interpretation is the possibility that the sectors are approaching a stage in which a financial bubble is about to happen. This interpretation is unlikely since only a few countries are utilizing this technology extensively and the businesses and entities involved in the related technologies are diverse and versatile.		
Financial stability	Finance	The interest-based financial potential indicates that the actual values have increased significantly over a period of twenty years for all the three variables by over one hundred percent in comparison to a future value as indicated in tables I, II, III, by AP_1, AP_2, and AP_3 respectively. The present value was calculated by taking the actual value of the first year, and multiplying by 2000 to get the cost of the 2GJ plant and then financing over twenty years with a 3% interest, using the average relative interest in the US for similar projects. The changes in the actual values over twenty years indicate the significant business financial interest in this industrial sector and its potential future worth.		
	Variations	The relatively low variance variations despite the large changes and disparities in the actual cost-values indicate that the changes are gradual and paced. The difference in value between the variances and the covariance is due to the large disparity between the actual values of the three industrial energies.		
	Risk Assessment 1. Portfolio risk estimated to be -17% 2. Portfolio risk is estimated to be -0.02% 3. Portfolio risk estimated to be -2%	The portfolio theory's risk assessment provides us with positive results. The direct solar energy desalination sector, despite its fluctuating data, is more stable in terms of its variation and the negative output is due to the data variations, and the fact that it is not a major sector, where there does not exist a single large scale desalination plant anywhere in the world. Some small mediocrely designed plants exist in the US and Australia. However, the gas-based desalination plants are a major factor in the seawater desalination industry. Although the risk associated with such sector is small, there exists very few gas-based large scale desalination plants and are almost all located in one area, Arabian peninsula, where gas is abundant and cheap. Gas is just like oil subject to political and economic stability of the gas producing countries, and or the areas where gas is produced. Such volatility could make such venture very costly outside the middle east, although the safety of such projects could be in-doubt in case of a war in the area. Not to mention that the gas-based turbines are complex and hard to maintain which may exacerbate the situation. Nuclear energy for desalination is becoming more popular due to the cost effectiveness of the nuclear energy on the long run and its larger risk assessment is due to the increasing variations in the cost of producing electricity through nuclear power, however nuclear power can be more cost-effective if utilized to produce both heat for desalination of seawater and to run electrical turbines for electrical energy. Although this sector is facing an expected snag, because of its diverse applications. Nevertheless, there is a new method for employing nuclear reactors that utilize radioactive elements with short-life cycles, and diminished capacity to make nuclear weapons.		

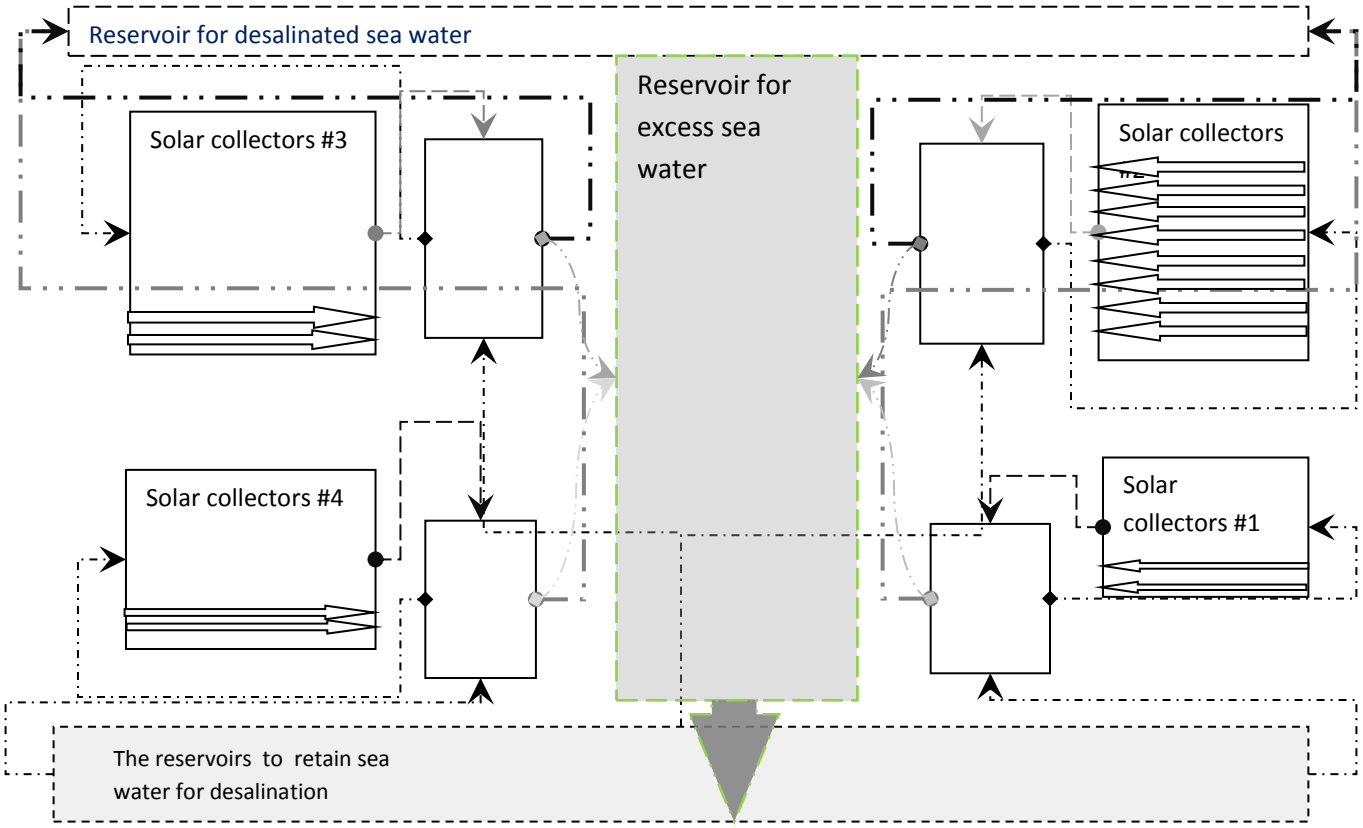
4. Conclusion:

Sadly, as it maybe, some sort of a hidden agenda to elicit large profits maybe in the works in the minds of many business executives looking to tap into this futuristically very profitable undertaking. Such mind-set is acceptable and encouraged, eventually mankind will benefit from such unsettling agenda. The analysis above demonstrates that seawater desalination is a viable business sector with a very profitable outlook. The utilization of direct solar energy is a definite plus technically and economically. The low cost of employing solar energy signifies its potential value. Although, land cost was never mentioned in the costing of this example, due to the government control over public land in Saudi Arabia, however, public land throughout the world can be leased at reasonable rates when utilized for public benefits. The low construction and maintenance cost of the solar energy for desalination of seawater outweighs legal, legislative or political obstacles. It is easy to build easy to maintain and unlike the other two options, is not subject to market or political fluctuation and wrangling, respectively.

5. References:

1. Odell Stephen Joseph. "Comparative Assessment of Coal-Fired and Nuclear Power Plants". Rensselaer Polytechnic Institute, Hartford, CT, December 2011.
2. Hogue. Michael T. "A Review of the Costs of Nuclear Power Generation". Bureau of Economic and Business Research, David Eccles School of Business, University of Utah, February 2012.
3. Al-Fulaij Hala Faisal. "Dynamic Modeling of Multi Stage Flash (MSF) Desalination Plant ". Thesis submitted for the degree of Doctor of Philosophy, Department of Chemical Engineering, University College London (UCL), July 2011.
4. Kaplan. Stan. "Power Plants: Characteristics and Costs ". CRS report for Congress, November 13, 2008.
5. Ashry Mohammed H. "Income and Employment in Single Sector Economies: Growth Through Diversification in Saudi Arabia". Department of Engineering Management and Systems Engineering, The George Washington University, 2008.
6. Díaz-Caneja José, Fariñas, Manuel. "Cost Estimation Briefing for Large Seawater Reverse Osmosis Facilities in Spain". RIDESA, Ramón Rubial nº 2, 48950 Erandio, Spai; Email: jose.diaz_c@pridesa.com.
7. Ashry Mohammed H. "A Large Scale Desalination of Sea Water by Solar Energy Uusing an Unconventional Seawater Collectors Scheme". CSREA Press, 2013.
8. Khamis Ibrahim. "Overview of nuclear desalination technologies & costs". Department Nuclear Energy, Division Nuclear power, IAEA International Atomic Energy Agency.
9. Squalli Jay. "Working Paper No. 05-01:Are the UAE Financial Markets Efficient?". EPRU, Zayed University, October 2005.
10. Torben G. Andersen, Tim Bollersev, and Ashish Das. "Variance-ratio Statistics and High-frequency Data: Testing for Changes in Intraday Volatility Patterns". The Journal of Finance • Vol. LVI, NO. 1, February 2001.
11. Harry M. Markowitz. "Portfolio Selection, Efficient Diversification of Investments". Newhaven and London, Yale University press, 1959.
12. James Barth, John Kraft, and Philip Wiest. " Portfolio Theoretic Approach to Industrial Diversification and Regional Employment". Journal of Regional Science, Vol. 15, No. 1, 1975.
13. http://www.economywatch.com/economic-statistics/country/Saudi-Arabia/ http://world-economic-outlook.findthedata.org/l/4708/Saudi-Arabia

(Figure I): the tide-based seawater desalination system (Ashry, Mohammed H; Csrea Press, 2013)



SESSION

POSTER PAPERS

Chair(s)

TBA

Reference Integrator: a workflow for similarity driven multi-sources publication merging.

Suman Reddy Mallipeddi, Carol M. Lushbough, Etienne Z. Gnimpieba
Computer Science Department, University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA,
{SumanReddy.Mallipedd; Carol.Lushbough; Etienne.gnimpieba}@usd.edu

Corresponding author: Etienne.gnimpieba@usd.edu, +1 605 223 0383.

~o~

Abstract: Making research sharable and reproducible faces big challenges. There are several sources available to manipulate personal, grouped, project and team publications for researcher work (e.g. NCBI, Arxiv, Google Scholar etc.). Duplicating publication becomes hard to manage. In order to evaluate the capability of a particular research, we have to collect and merge all publications from different sources. There are no tools to collect the researcher publications into a single document. As a solution, we are introducing a new web tool, Reference integrator that helps researchers to select their publications from different sources and merge them into a single document and save the document in Pdf or Doc format. It extracts the publications from different sources (Publications.li, NCBI, DBLP, Google Scholar, Arxiv, Mendeley, and Microsoft Research), removes the duplicate publications and delivers the list of publications in a single document. This saves valuable time for the researcher by automatically providing the list of all publications in one document that can be download (pdf, text) or shared with others.

Availability: Reference Integrator is freely available at http://jacksons.usd.edu/ReferenceIntegrater/ .

Keywords: Sharable and reproducible research, merged publication**,** publications extraction, publication similarity, matched publication.

1.Introduction

Web 3.0 development forces researchers, in order to be visible, to create multiple workspaces for their work. A given researcher in bioinformatics for example can manage over 100 workspaces online (google citation link, Mendeley, Endnote, NCBI, etc.). The key problem consists in extracting at a given moment the merged list of all publications produced by the researcher or in a given project. In addition the research reproducibility challenge depends on the ability to easily share the output documentation and results [1].

This work helps the researcher to select all the publications and combining them in a single document to avoid multiple publications. This work also provides a matching rate of the publications before generating the merged result (default matching percentage: 70)[2]. The user can change the matching percentage by moving right on the slide bar (which increases the matching percentage) or to the left side (which reduces the matching percentage) and produce the result accordingly.

2. Methodology:

Figure 1. Presents the proposed 4 steps workflow use to provide the merged publication list. Reference Integrator extract publications from different sources and merges them into a single document, eliminating the duplicate publications collected from different sources. This is done with four main features.

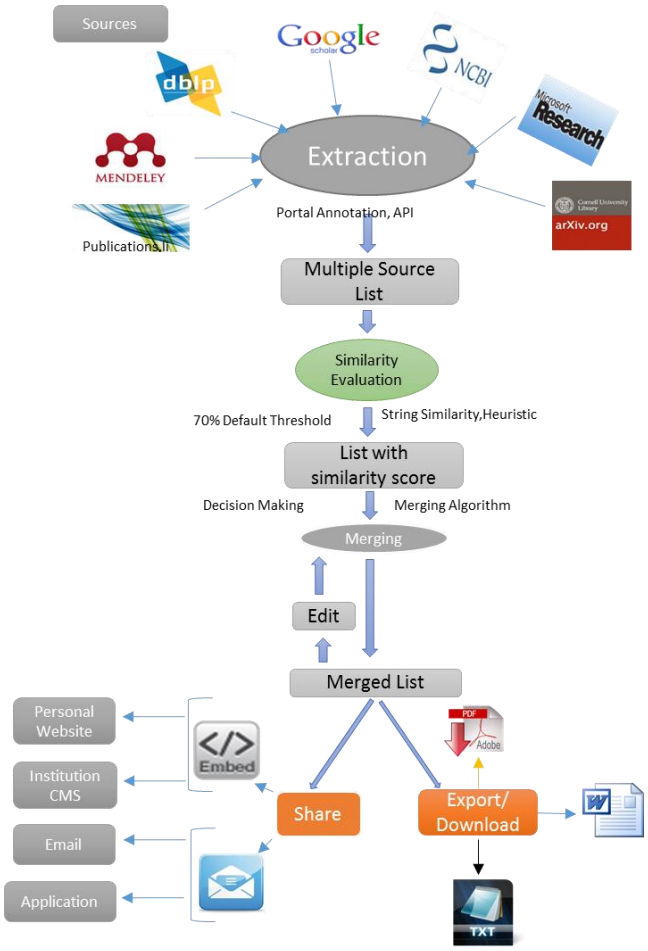


Figure 1: Reference Integrator Global workflow

a) **Extraction:** Multiple sources are selected and submitted for extraction. Using Portal Annotation and API, the required publications are extracted.
b) **Similarity Evaluation:** Using a string similarity evaluation algorithm, we checked the similarity of the listed publications individually (by default threshold: 70%) and get the list of publications with related similarity score.
c) **Merging:** applying merging and decision making algorithm for the list of publications gives the merged publications, eliminating duplicated publications.
d) **Edition and archive**: From the obtained merged list user can edit the merged list and download in three different formats (PDF, DOC, and TXT). User can also share the result using email or by copying the embedded source code of the merged publications (which helps to customize the personal website).

4. Conclusion:

Reference Integrator is the first consistent workflow for managing publication data from different data sources. The similarity algorithm allows customization of the merging process by assigning appropriate score, reduces repetitions and supports sharability. A researcher can extract publications from the sources where the publications are stored and check for similarity and store them in a single document. Integration of this process in a workflow allows us to save considerable time. This work constitutes an important step in the reproducible research challenge.

Funding: This work was made possible by SD-INBRE Grant #P20RR016479-09 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

References

[1] J. P. Mesirov, "Computer science. Accessible reproducible research.," *Science*, vol. 327, no. 5964, pp. 415–6, Jan. 2010.

[2] M. Mednis and M. K. Aurich, "Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models," *Biosyst. Inf. Technol.*, vol. 1, no. 1, pp. 14–18, 2012.

SESSION

LATE BREAKING PAPERS AND POSITION PAPERS: INFORMATION AND KNOWLEDGE ENGINEERING

Chair(s)

TBA

Operational and Organizational Dimensions of the Bid Process Information System (B.P.I.S.)

Sahbi Zahaf
Higher Institute of Computer and Multimedia
MIRACL Laboratory, Sfax University, Tunisia
sahbi@zahaf.net

Faiez Gargouri
Higher Institute of Computer and Multimedia
MIRACL Laboratory, Sfax University, Tunisia
faiez.gargouri@isimsf.rnu.tn

***Abstract*— Bid process translates the techno-economic expertise, which partners build in a cooperative way. It is a key business process which evaluates the results of different trade tasks: hence, it influences the company's survival and strategic orientations. Therefore, the Information System that supports this process must be characterized by integrity, flexibility and interoperability. Nevertheless, the urbanization approach, on which we rely to implement this system, has to deal with "three fit" problems. To overcome these problems, we suggest addressing these exigencies following an operational dimension which remains responsive to other dimensions: the organizational and decision-making ones. However, the cooperative dimension covers the remaining dimensions. In fact, it ensures the consistency and the interaction between the different dimensions. We are interested in identifying the tools and applications needed to achieve a bid process. Specifically, we are focused on the characteristics of operational and organizational dimensions.**

Keywords- Bid process, Information System, ERP, Ontology, Ontology Design Pattern, Organizational Memory.

I. INTRODUCTION

The owner calls for a bid to benefit from a product or service; after some period, he receives many proposals from different participating companies which submit their responses to this call for tenders (bid process). A bid process embodies a techno-economic proposal (a technical expertise backed by a financial offer). Such a contribution translates preliminary, the recommendations proposed by each contributor, either to reconstruct the desired product, or to organize the required service. Therefore, it is an elementary study that takes place before negotiating the contract with the owner, i.e. before launching the project. The bidder (company that pilot the bid process) might appeal to some partners, especially during the construction of the technical proposal. In the meantime, we can have new calls for bid following the first one and so on. Each participant seeks to cover and meet mainly his financial benefits because it is recognized as the most powerful and incorporating dimension. It interrelates with and touches the other dimensions like the: technical, political, social, environmental, etc. In each evolutionary step of the bid process, collaborators can renounce their participation when they recognize that such a deal does not cover their objectives. Indeed, most of the participants who decide to pass their last proposal, make a thorough evaluation exercise that tests the feasibility and effectiveness of the offer, through the technical, economic, temporal, and risk indicators. Thus, the bid process is a key business process of the company insofar as it directly affects its future and locates its expertise as well as its competitiveness in relation to its competitor [1]. The owner focuses on the best offer, which meets its requirements and which covers the eminent interests. Once gained the offer, the bidder and his collaborators create the bid process team. The latter makes part of the industrial consortium which targets to realize the project in practice.

Certainly, it is the company's agility and competence that allows acquiring the customer's confidence and interest, and as a consequence wins the offer. It is to remember that doing business means taking risks, something that can influence the company's growth and survival. So, the strategic management of business is a major and current concern to an innovative enterprise. The latter promotes to restructure its Information System (I.S) around its trades and business processes.

The I.S is the executing support of the processes of business. It directly influences the internal and external environmental requirements of the company. Internally, the consistency of an I.S depends utterly on its degree of integrity, flexibility and its internal interoperability. While externally, I.S agility always depends on its flexible capacities and external interoperability. Nevertheless, the reality shows agility and consistency problems on both inter as well as intra levels. We prove this premise when we apply the urbanization approach [7]. It consists of three transposition problems dubbed as "*three fit*": (i) "*vertical fit*" (lack of integrity and lack of extensibility); (ii) "*horizontal fit*" (lack of flexibility and lack of internal interoperability); (iii) "*transversal fit*" (lack of openness and lack of external interoperability).

Our main objective is to suggest I.S which supports the bid process (B.P.I.S.). Actually, this system must be: *integrated*, *flexible* and *interoperable*. We treat these aspects following four dimensions: (i) the *operational dimension* which aim to specify the bid operation process designated for a specific project; (ii) the *organizational dimension* which targets to specify the set of skills and knowledge that the company had acquired due to previous bids in order to eventually reutilize this patrimony in future bid projects; (iii) the *decision-making dimension* whose goal is to optimize decisions on the market's supplies; and finally, (iv) the *cooperative dimension* which

seeks to plan the inter-enterprises cooperation that take place during the construction of the techno-economic proposal of the offer. In this work, we are interested on the operational and organizational dimensions.

This article is organized as follows. First, we identify the required characteristics I.S supporting a bid process, and we focus on the hindrances encountered while implementing I.S: problems of "*three fit*". Then, we suggest an urbanized B.P.I.S, to overcome these limitations. After that, we specify successively the operational and organizational dimensions of a bid process I.S. We finish this work by a conclusion that opens the horizons to our future research projects.

II. REQUIRED CHARACTERISTICS AND PROBLEMS IMPLANTATION OF THE BID PROCESS INFORMATION SYSTEM

A. Required Characteristics of the Bid Process I.S

The I.S which supports a bid process must be:

Integrated: the company regularly participates in bids, during which, it reutilize its technical skills and business, to subsequently take optimal business decisions. It is in this way that its I.S must be integrated, i.e. capable of creating a comprehensive synergy including (hardware, software, features, and users), so that they cooperate within a single homogeneous system. Integration is a way that ensures consistency and harmony on different levels:

- *Business level*: the bid is a business process focused on the operating results of the business processes. The company is not integrated at this level only if it manages to standardize processes, so as to create coherence at the level of applications which feed the appropriate functions.
- *Informational level*: I.S must structure the knowledge and skills acquired during past bid experiences. The objective is that the enterprise can exchange and share its patrimony while contributing to new offers. The company is not integrated at this level, only if it manages to give a uniform image on its own capital of object that it manipulates in its functional scope and allows its reutilization from one action to another.
- *Decision-making level*: the company should regularly take strategic decisions to each evolutionary step of the bid project. Thus, its I.S must be able to optimize business decisions, while having the ability to predict the feasibility of the offer on the basis of its targeted objectives. The company is not qualified as expert at this level, only if it can decide easily on its own capital of object which it adapts from one action to another.
- *Technical level*: the applications needed to achieve a bid must cooperate appropriately in order to exchange the right information at the right time, which allows the user to take the right decision. This premise requires that the semantics undertaken by the exchanged data be interpreted in the same way by all enterprise applications.

Flexible: the company should survive in an unpredictable business environment: each bid process is realized in a specific context and consider specific solutions. That's why, I.S needs: on the one hand, to be able to overcome the market changes, from one offer to another (adaptability, scalability and extensibility) and on the other hand, to be able to react, on the right time, with the business agility (competitiveness and responsiveness). The company is considered flexible, when it manages to operate, adapt and easily extend its resources at different levels (business, informational, decision-making and technical) to fill and quickly meet the offers' terms during the different occasions of business that it accomplishes.

Interoperable: the company should not address offers and should not conduct business autonomously, and that's why I.S must be interoperable both inside, and outside its functional scope. Obviously, the interoperability is the fact that several systems can operate together while preserving their heterogeneity and autonomy. Thus, internal interoperability is a prerequisite to build an integrated I.S. On the other hand, given the competitive environment in which the business is involved, such a system must foster intercompany cooperation on the spot and in a dynamic way, whenever it is necessary to organize a partnership relation of a bid project, especially, during the construction of the techno-economic proposal: so it requires planning the dynamic interoperability at the external level. Practically, the enterprise is interoperable, unless its I.S is integrated internally and it manages to plan, to synchronize and to exchange: its trades, its data, its resources and its processes, easily with partners from the outer world, and this should happen despite their semantic differences.

B. Implantation of the Bid Process I.S: "three fit" problems

We rely on the urbanization approach to establish an I.S dedicated to the exploitation of the bid process [7].

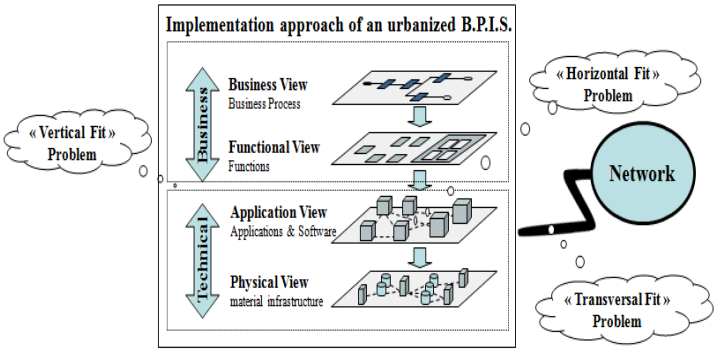


Figure 1. Urban I.S reference model: problems of "three fit" [7].

This approach is described according to four levels: (i) business view (it represents the company's operations and business processes that are necessary for the bid); (ii) functional view (it represents the functions and flow information towards business processes regardless of the technologies used); (iii) application view (to conceive all of the applications used by stakeholders to support functions and

flows, and also to equip the process); and (iv) physical view (it is the infrastructure on which are implemented the application blocks to support the technical architecture). Nevertheless, this approach is facing the following problems (Fig. 1):

"Vertical fit" problem: the business and functional views describe the trades' needs. They are abstract. But, the application and physical views represent definite implementations. It would be difficult to gather the data that allow scrutinizing the company's operations. In fact, this results a fragmentary data in work system and reduces the efficiency of the company. These circumstances prevent from having a system which gives a complete and integrated image on company's inner environment. Integrity, scalability, consistency and transposition are *"vertical fit*" issues that extends from a business infrastructure (logic) to a technical infrastructure (physical) in the company's I.S.

"Horizontal fit" problem: the "*horizontal fit*" translates not only the applications' problems of identification (induced by the "*vertical fit*" problems) that cover the entire infrastructure of the company's business, but the intra-applicative communication problems (internal interoperability) to ensure the interactions between applications of the same technical infrastructure in the company (the same homogeneous system). Such failures make the global system disintegrated and little evolutionary.

"Transversal fit" problem: the "*transversal fit*" translates the problems of inter-applicative communications (external interoperability carried out dynamically).

To conclude, the "*three fit*" problems prevent the transversal exploitation of bid process, and as a consequence, it is difficult to have a unified bid vision. Solving such problems needs to meet the requirements associated with integrity, flexibility and interoperability on internal as well on external of I.S. The internal interoperability concerns the applications within the same enterprise. But, the external must be dynamic, on demand, in order to realize a common goal within a virtual company. It should be noted that an I.S is flexible if it is: (i) extendable (relies on a technical architecture that promotes its branching); (ii) evolving (able to withstand a large amount of treatment without affecting its architecture); and (iii) adaptable (promotes reutilization which is based on the specification of I.S invariants). Thus, the flexibility qualification is easier to set up and realize if the I.S is integrated. However, an I.S is considered integrated only if: on the one hand, the coherence between all the applications is ensured (business and technical levels), and on the other hand, the uniqueness, relevance and reliability of information that feeds these applications is guaranteed (informational and decision-making levels). Still, this integrity assumption is valid only if intra-applicative communications are carried out without ambiguity. Henceforth, interoperability is a necessary condition for I.S to describe it as integrated. Finally, to get a flexible I.S, it must be integrated. However, this last feature is ensured if we solve the interoperability requirements. Indeed, this problem arises when I.S integration of the company (internal interoperability) is concerned as well as when a virtual company is organized (external or dynamic interoperability). However, the non-satisfaction of the interoperability requirement incurs significant costs associated primarily to time and resources which are presented to develop exchange information interfaces and knowledge sharing (technical interoperability), following a common semantics (semantic interoperability) and which are supposed to train actors and adapt organizational procedures (organizational interoperability). Such assumptions influence negatively the overall performance of enterprises and more precisely the costs as well as the deadlines of getting the services expected while realizing a bid process.

III. Our Proposition: The Urbanized B.P.I.S

Our objective is to set up B.P.I.S which is *integrated*, *flexible* and *interoperable*.. We are interested, not only in implementing the right tools to achieve bid in one homogeneous system (integrated), but also, in solving the problems related to interactions intra, and even inter-applicative (interoperability). Our aim is to be able to exploit this system in different bids (flexibility). So far, we dealt with "*three fit*" problems.

In this context, we suggest to meet these requirements by relying on four essential dimensions. They are: (i) the *operational dimension* that serves to specify the bid exploitation process by undertaking a specific project; (ii) the *organizational dimension* which allow to organize the set of skills and knowledge, that the company acquired during the previous bid in which it participated: the objective is a possible reutilization of this patrimony in future bid projects; (iii) the *decision-making* dimension which aims at optimizing and making the right decisions that concerns the market offers and that takes place during the company's eventual participation in bid processes; and (iv) the *cooperative dimension* which aims at ensuring communication intra-enterprise (internal interoperability) and at planning the inter-enterprise communication on demand , in order to realize a common goal (dynamic interoperability). For example, while creating the offer's technical proposal, one needs to organize inter-enterprise collaborations.

Afterwards, we treat respectively: *flexibility* through the *operational dimension*, *integrity* through the *organizational* as well as *decision-making* dimensions; and *interoperability* through the *cooperative dimension*. In fact, as far as the company is concerned, a flexible bid exploitation (*operational dimension*) requires an eventual integration at the level of I.S.; this fact targets reuse the best skills (*organizational dimension*) and adapt the best decisions (*decision-making dimension*), and learn from past bid experiences. All of this requires an internal interoperability within the participating company in the bid process to seek a homogeneous and coherent form of its I.S (*internal cooperative dimension*); and dynamic interoperability, at the level of a virtual company, built to realize in common a bid project, the fact that allows a

better coordination and collaboration between various involved stakeholders (*dynamic collaborative dimension*). However, it is the *operational dimension*, which is deemed to be the main focus of our system, which acts and remains sensitive to variations in all other dimensions: the *organizational* and *decision-making* ones. These various dimensions are covered by the *cooperative dimension* itself (Fig. 2). Indeed, the *operational dimension* takes the *organizational dimension* as a basis in a specific bid. This premise is justified by the fact that the company reuses its own capital of objects for different actions, depending on its needs. This pushes the company create more and more products whose life cycle is shorter than those made in the past. This hypothesis is based on the re-design of existing products or creating similar design products, rather than, on producing new ones. In another way, the *organizational dimension* is based on the *operational one*, both while constructing a bid proposal, or after finishing it. In fact, the former (bid proposal in progress) comes to readjust the proposal, while the latter (bid proposal is finished), comes to update the company's capital through integrating the set of knowledge and skills built during this project. It is worthy to note that this assumption is beneficial for the company's maturity, even when the company abandons its participation in the offer (the company closes its participation before the completion of the bid). In all cases, even if the owner chooses the proposals of other companies (the bid proposal is unsuccessful), a possible updating can take place and influence positively and beneficially, at least for future bids. If we follow the same logic, the *decision-making dimension* relies on the *operational dimension* and vice versa, which may or may not consolidate the company's participation in the bid. It becomes evident that the *decision-making dimension* is inexorably related to the *organizational dimension*. All these dimensions are based on the *cooperative dimension*: on the one hand, the realization of the bid can be cooperative (*operational dimension* relies on the *cooperative dimension* for collaborative planning and for the creation of a product); and on the other hand, decisions can be cooperative (the *decision-making dimension* relies on the *cooperative dimension* for a collaborative decision during a definite bid).

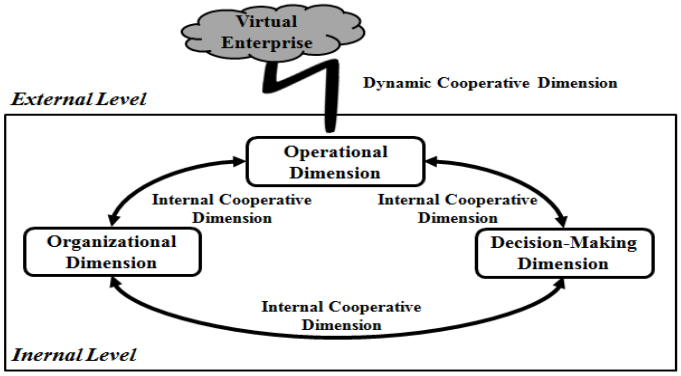


Figure 2. Our Bid Process Information System (B.P.I.S.) dimensions.

We suggest exploiting these different dimensions, while relying on six main approaches, and by describing our urbanization approach of the I.S. Indeed:

The implementation of the Lean Manufacturing [15] approach allows designing a product perfectly adapted to the needs of its client and this product can be cheaper in costs but not less efficient than its expected services. This approach integration adds value to the technical proposal construction which materializes a bid (Lean Manufacturing participates in solving the problem of "*vertical fit*"). The BPM (Business Process Management) incorporation [16] allows to model business and skill processes, particularly, the bid process. This approach facilitates the alignment of an integrated I.S with strategic directions, regardless technological constraints (BPM participates in the resolution of the problem of "*vertical fit*").

The KM (Knowledge Management) involvement [6] allows formalizing and modelling bid knowledge, whether explicit or implicit, in order to make them operational by the company during different bid projects that it realizes. This approach facilitates establishing a language and covers implementing the organizational dimension within the company (KM participates in solving the problem of "*horizontal fit*"). Furthermore, KM permits improving and responding to individual, collective and organizational learning acquired during a bid process. This hypothesis suits perfectly a bid context (KM takes part also in resolving the problem of "*transversal fit*").

The BI (Business Intelligence) integration [17] allows relying on methods, in order to provide decision-making assistance to those involved in a bid process. Therefore, this approach facilitates implementing explicitly the decision-making dimension within the company (BI participates in the problem resolution of "*horizontal fit*").

The SOA (Service-Oriented Architecture) [2] helps developing an easily flexible, extensible and adaptable I.S which can be materialized by a set of reusable application components. These application blocks communicate the practical implementation of "services" (clearly defined function in a way that makes it independent of the technical platform). The SOA facilitates the communications' standardization, intra-applicative (SOA participates in the resolution of the problem of "*horizontal fit*"), as well as inter-applicative (SOA participates in solving the problem of "*transversal fit*"). SOA offers an innovative solution to manage the interface between the business needs and its technical implementation (SOA participates in the resolution of "*vertical fit*" problem).

However, the company that takes part in a bid process needs to exploit its "services" at a distance, to promote collaborative work with its partners, such as the work constructed during the technical solution. The integration of Cloud Computing [9], allows the enterprise data and applications to be accessible and usable via the internet (Cloud Computing participates in the resolution of "*transversal fit*" problem).

We can deduce that (Table 1): Lean Manufacturing, BPM and SOA allow us to overcome "*vertical fit*" problems and thus cover all the dimensions defined previously. KM, BI and SOA allow us to overcome "*horizontal fit*" problems as follows: KM permit to cover the *organizational dimension*, BI can cover the *decision-making dimension*, and SOA enable to cover the *cooperative dimension* at the level of the enterprise. In addition, SOA and Cloud Computing permit to overcome

"*transversal fit*" problems and hence cover the *cooperative dimension* to the level of a virtual company. Also, KM participates in solving the "*transversal fit*" problems and consequently cover the *organizational* and *decision-making dimensions*. To this end, it is the *operational dimension* which can be supported by these six approaches. In other words, the satisfaction of this assumption enables us to have a flexible, integrated, and interoperable, I.S something that assures us a better exploitation of the bid process.

TABLE I. BID PROCESS I.S IN THE CORE OF THE COUPLING CAPACITY OF THE SIX APPROACHES.

		Flexibility		
		Integrity		Dynamic Interoperabily
		Internal Interoperability		
		Vertical Fit	Horizontal Fit	Transversal Fit
Operational Dimension	Organizational Dimension	Lean Manufacturing / SOA/ BPM	KM	KM
	Decision-Making Dimension		BI	
	Cooperative Dimension		SOA	SOA / Cloud Computing

We suggest filling these dimensions while relying on the following hypothesis (Fig. 3): "ERP (Enterprise Resources Planning) [10] allows us to build the techno-economic proposal of an offer (cover the *operational dimension*), as we rely on the organizational memory (to cover the *organizational dimension*). The set of solutions that make this proposal realistic, are going to be evaluated by a Data Warehouse (to cover the *decision-making dimension*)".

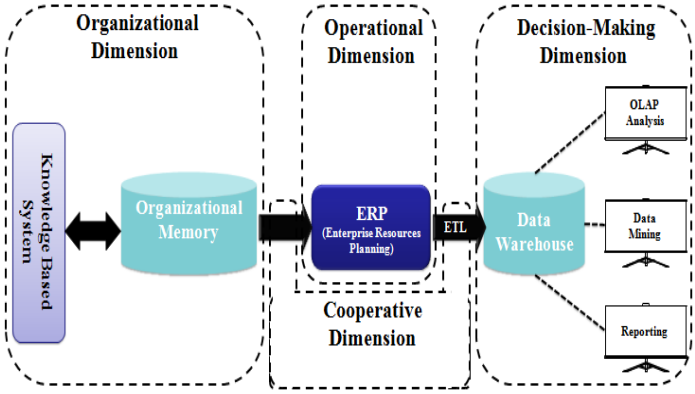


Figure 3. Our urbanized Bid Process Information System (B.P.I.S.)

IV. OPERATIONAL DIMENSION OF THE URBANIZED BID PROCESS INFORMATION SYSTEM

The ERP [10] aims to pilot the enterprise processes. In this framework, ERP producers suggest integrated packages based on the common skills between enterprises such as: financial, purchase and sales management, production management, technical data management (items, nomenclature, resources, manufacturing process), logistics management, etc. More precisely, these software producers build integrated application modules implementing the market's best practices for each function. Thus, the ERP is the most suitable for the complex processes exploitation within the enterprise, particularly; it helps to exploit the bid process. In fact, the bid process relies on the previously enumerated functions for its exploitation, notably; within the context of making the bid's techno-economic proposal (the ERP covers the bid's *operational dimension*). Nevertheless, there are primordial functions for the bid process evolution, but these functions are not treated by the ERP such as: expertise management (the ERP does not cover the *organizational dimension*), risk management and decision-making management (the ERP does not cover the *decision-making dimension*). This justifies our choice, in the previous paragraph which states to support the ERP by other tools in a way that enables treating a bid process. In fact, the ERP targets to meet the bid process *operational dimension.*

The success of implementing an ERP project dedicated to a particular enterprise must pass by design and platform coupling of reengineering and integration, internal versus external, trades and business processes. We are facing an exigency related to a specific governing strategy of each enterprise on each trade. This strategy intends to establish a simultaneous correspondence and alignment between two distinct objectives: the first refers to the ERP agility (the ability to adapt to different management modes and to harmonize its exchanges with the appropriate partners). However, the second objective is linked to management processes, particularly, those exploited by a bid process (the enterprise's ability to innovate in an incremental and dynamic way opens the ground for a better management of costs, deadlines, risks, etc.).

Our goal is to set up an ERP ad-hoc and its strategy as a solution. Therefore, the technical capacities and management coupling impose themselves. However, adaptability to the context or "conduit change" is a major factor of success for an ERP project throughout a functional integration. It is, on the one hand, a "governing affair" conduit which assures innovation in the management of business and trade processes. On the other hand, it is a "governing technology" conduit which ensures the dynamics of these processes. Besides, studies formalizing the combination between the two previously mentioned objectives are rare or non-existent. Indeed, it is not easy to deploy a solution that aims to respond to such governance, combining an evolutionary and incremental approach to implement a technology that seeks the company's I.S. agility while allowing a better management of its business and trade processes. As a result, we rely on the NICT (new Information and Communication Technologies) to answer this problematic. In fact, ERP has some limitations on management and technical levels. As a consequence, since the ERP relies on NICT, we are able to meet the limits detected on those two levels. Certainly, this solution enables us a better management of the bid process.

It is true that the ERP aims to manage the bid process and its information flow. However, the fact that the ERP relies on a BPM approach, this allows to model the bid process and to subsequently align the ERP on strategic directions, without undergoing many technological constraints. Such a strategy helps the bid process standardization (this allows to have different ERP(s) with different layers of trade standards while realizing a bid process). Hence, the cooperation between enterprises and more accurately between ERP(s) (let us remember that an ERP covers a bid process *operational*

dimension) is easier to realize, and in this way we get the impression that we are working on the same ERP (this solution is highly recommended during the collaborative construction of the offer's technical solution). Consequently, the BPM will help us to construct an ERP in a modular way by assembling trade components weakly coupled. It remains only to solve the communication between different technical infrastructures which materialize the different ERP(s). We previously showed that the SOA allows solving such problems.

The ERP enables to align processes best practices. However, it is unable to follow these good practices. Thus, the ERP does not adapt with the continuous improvement of the enterprise in its affairs. The fact that the ERP relies on Lean Manufacturing strategy enables to overcome this limitation and helps the enterprise to establish a culture and a permanent maturity that revolves around good practices during the different contributions in different bid processes.

The ERP does not permit reusing the enterprise's expertise acquired due to previous bid processes. As a result, the ERP must rely on an external system that enables to collect, formalize, restitute the data and skills, in order to make them available, operational, and exploitable by an ERP during a specific bid process, in real-time. Therefore, the fact that the company relies on a KM approach and more accurately on an OM this enables the ERP to effectively exploit the enterprise's internal language in different bid processes.

The ERP is designed to collect the event traces, but it is not created to help the decision-making process. Consequently, the ERP inability to treat uncertainties and unexpected events limit its use to support the decision-making process in an environment of dynamic production. Therefore, the fact that the ERP is based on BI approach, and more precisely on a DW, enables the enterprise's employees to get a rapid and synthetic access to strategic information. Furthermore, these employees can easily adapt their decisions taken in a past project to a specific bid process project.

We can deduce that a coupling between an ERP and management capacities (BPM, Lean Manufacturing, KM et BI) are highly recommended to meet the "governing affairs" level of a enterprise in a bid process project. Following the same logic, we demonstrate in what follows that a coupling between ERP architecture and technological capacities (SOA and Cloud Computing) is absolutely necessary to meet the "technological governing" level of the company in a bid process project. The structures provided by an SOA allow implementing ERP ad-hoc architecture in the enterprise hence helping it to surmount environmental turbulences, to respond to agility affairs quickly, and to improve the management of its markets and its bid processes. Moreover, ERP solution based on SOA architecture helps to develop, at the very heart of the enterprise, new skills thanks to reutilization. This strategy enables to enhance the company's expertise and to establish, as a consequence, a permanent culture of change. This is revealed to be necessary for the company's eventual adjustment in different contexts. These characteristics strongly favor the management by processes (this is strongly recommended to be able to manage and exploit the bid process). In fact, the architectures suggested by SOA are flexible, adaptable reusable and extensible to a great extent. These architectures enable to easily integrate the new affairs costs realized by the enterprise in a specific bid process and to constantly develop trade and business processes modeling. Furthermore, the ERP deployed with SOA architecture ensures a low intra-applicative coupling (low coupling between ERP and the other applications used in the enterprise) as well as a low inter-applicative (low coupling between different ERP of different enterprises to realize a bid process).

In certain cases, the ERP needs to exploit distant functions which are hosted externally. This requirement is realized if the enterprise includes Cloud Computing approach during the implementation of its applications. Thus, the ERP must interact with applications of SaaS (Software as a Service) [3] mode each time it needs to exploit new functions which it does not cover. It is noteworthy that this solution is strongly recommended during the collaborative construction of a bid process proposal.

Finally, we validate the hypothesis that we departed from: *"the fact that the operational dimension ERP is based on six approaches (BPM, Lean Manufacturing, OM, BI, SOA and Cloud Computing), this helps us to meet other dimensions (organizational, decision-making and cooperative). Such solution allows us to have a flexible, integrated, and interoperable I.S., which helps us to ensure a better exploitation of the bid process."*

V. Organizational Dimension of the Urbanized Bid Process Information System

To work on a specific bid implies the intervention of several collaborators. Certainly, these contributors exchange knowledge and information flows. However, its environmental differences lead to various representations and interpretations of knowledge, and therefore, on the same corpus, different skills and semantics overlap (interoperability problems). Some occurring failures are described in terms of a set of five conflicts [13]. Practically, the *syntactic conflicts* are the results of different terminologies used by stakeholders on a same application domain. *Structural conflicts* are related to different levels of abstraction which aim at classifying knowledge within a virtual company (bid team). *Semantic conflicts* concern the ambiguity that emerges due to the stakeholders' reasoning in the development of the technical and economic proposal. *Heterogeneities conflicts* are due to the diverse data sources (specifications, owner, experts, collaborators, etc.). Finally, *contextual conflicts*, come mainly from environmental scalability problems, and in fact stakeholders can evolve in different environments.

In order to answer to these various conflicts, we suggested in [18] an OM sustained by an ontological framework in order to operate on certain business processes, we can accommodate this to realize a context for a bid process. However, this memory needs to be empowered by a knowledge-based system, to operate, share and automatically reason on business knowledge between different stakeholders. This system allows

overcoming the *structural* and *syntactic conflicts*, and as a result it solves the problem related to knowledge acquisition.

In return, it does not solve the ambiguities related to knowledge representation (*semantic* and *contextual conflicts*). In the perspectives to answer the requirements related to solving the *semantic* and *contextual conflicts*, we suggest an ontological modelling framework of business knowledge.

Our approach which seeks to construct an ontological framework for the business operation process is jointly supported, on the one hand, on the specialization of the founding ontology DOLCE [12] which apply the method OntoSpec [11], and on the other hand, on the Ontology Design Patterns (ODP) [8] relating to kernel ontologies:

- A specialized founding ontology DOLCE allowed us to master the complexity of conceptual modelling. Hence, it solves problems related to semantic conflicts. Accordingly, we reutilized concepts from DOLCE to specify generic concepts related to business processes (DOLCE is the backbone of the OntoSpec method). Also, this work allowed us to achieve a modelling of different levels of abstraction.

- Ontology Design Pattern (ODP), allowed us to master the complexity of consensual modelling at the generic level, this solution solves problems related to *contextual conflicts*. Indeed, the use of these ODP is based on the reutilization of the ontological modules already designed and evaluated in other areas [8]. It is worthy to note that the concepts used in ODP are defined according to concepts and relations issued by the specialized ontology DOLCE. Practically, we defined the *ODP relating to a business process treatment*, *ODP resources*, *ODP risks*, *contextual ODP*, and *ODP construction products* [18].

The application of our proposal, framed in a context of identification of bid knowledge on a specific project, aims to define four types of ontologies: (i) foundational ontology (specialized DOLCE which defines the invariant concepts of business process); (ii) kernel ontology (ODP for the reutilization of the invariant concepts of business processes); (iii) domain ontology (specialized in concepts relating to the kernel ontology in the bid domain); and (iv) application ontology (specialized in concepts of the bid ontology domain in a particular application: bid in a specific project).

The produced business skills must be stored for possible future use. For this reason, we suggested an OM for the management of business processes (Fig. 4). We use this memory to exploit the bid process in the context of a particular application. In fact, this OM can deal with the problem related to the capitalization and the restitution of knowledge, and therefore, it resolves *conflicts of heterogeneity*. We organized our OM with a set of five sub-Memories: a *reusable resources memory*, a *context memory*, a *roles memory*, an *action memory*, and *uses cases memory* [18]. These memories are supported by different ODPs enumerated above, and by the different models of CommonKADS [14].

In a specific bid project, our starting point is the tender issued by the owner (a tender is a set of specifications). Concretely, a tender defines and details the set of elements to take in order to execute and manage the project. Its objective is to describe explicitly the desired functionality for future product: owner vision. The analysis of specifications allows alimenting *context memory* (organization model, agent model, and ODP contextual) and also *action memory* (task model and application model realized in the form of application ontology). A well explicit context frames the implementation environment of different uses of reusable resources. Henceforth, the *memory of reusable resources* stores the knowledge generated by the set of objects and reusable concepts that the company manipulates and controls in its routine activities (ODP treatment process, ODP resources, ODP construction product and ODP risk). Thus, reusable resources participate in construct the techno-economic proposal of the offer while being based on *roles memory*, the latter stores the knowledge generated to describe the use of a reusable resource within a given context. *Cases uses memory* describes knowledge built for each bid proposal departing from the content of other sub-memories. Each bid proposal will be subsequently evaluated by indicators (technical, business, temporal, risks, etc.). Thus, for a particular solution we suggest to construct a Bid Memory which can be constituted by: a technical referential for product design, cost referential and price one to evaluate the case, and a risk referential to specify the possible risks associated with its design. Bid Memory (Fig. 4) will therefore be the dynamo for the bid exploitation, notably to prepare the proposal (OM stimulates the bid memory, which enables to cover the *organizational dimension*).

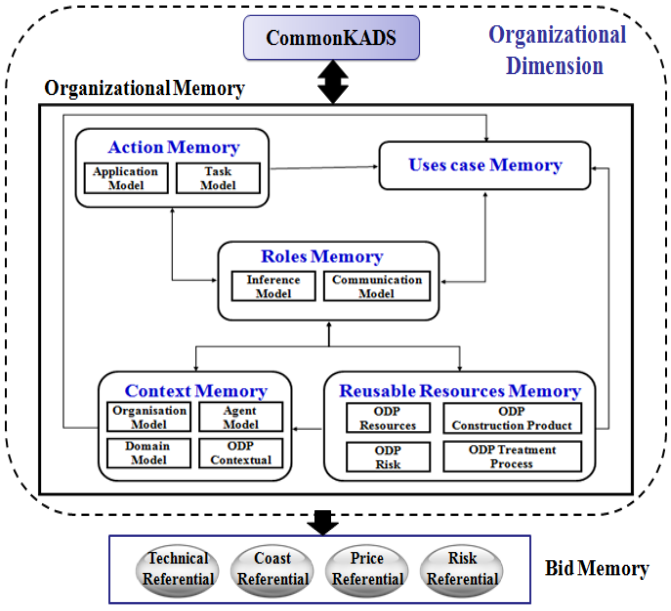


Figure 4. Our B.P.I.S.: Characteristics of the Organizational Dimension.

Henceforth, the ERP will empower these different referential during the construction of the techno-economic

solution of the offer (the ERP permits to cover the *operational dimension*). The set of solutions included in this proposal will be subsequently evaluated by the DW (DW can cover the *decision-making dimension*).

VI. CONCLUSION AND PERSPECTIVES

In this article, we presented our methodology of implementing a system of management for the exploitation of the bid process. Initially, we showed that such a system must be *integrated*, *flexible* and *interoperable*. However, during the implementation of this system, "*three fit*" problems (*vertical*, *horizontal* and *transversal fits*) fail the inclusion of such requirements. To overcome these deficiencies, we proposed to address the I.S design following four dimensions: the *operational dimension* (tackles *flexibility*), the *organizational* and *decision-making dimension* (tackles *integrity*) and the *cooperative dimension* (tackles *interoperability*). We are not only interested in implementing the right tools to achieve a bid process in one homogeneous system (*integrated*), but also, in solving problems related to interactions intra, or even inter-applicative (*interoperability*), which enable exploiting this system in different bids (*flexibility*). We have proposed to define bid memory for each project. Such a memory is composed of: a referential of products design, a cost and price referential to evaluate the case and a risk referential to estimate the risks related to its contribution. However, a bid memory can be alimented by an OM supported by an ontological framework for the exploitation of business processes (these two memories target the *organizational dimension*). Moreover, we showed that the ERP is the most suitable tool for the exploitation of a bid process, thus, it is a constructing support techno-economic solution of the offer (the ERP targets *operational dimension* of a bid). The set of solutions realizing this proposal will be subsequently evaluated by the DW (DW targets *decision-making dimension*). In order to assure reliable interactions, at both the internal and external levels of the company, we opted to deploy applications defined by our system which is of reusable architecture (SOA and Cloud Computing target the *cooperative dimension* at the level of a virtual company). Our work opens the horizon to exploit this suggestion within a concrete bid project.

REFERENCES

[1] Alquier, A.M, Tignol, M.H. : Management de risque et Intelligence Economique, l'approche PRIMA, Economica, Paris (2007).

[2] Bean, J.: SOA and Web Services Interface Design, Principles, Techniques and Standards, Elsevier, Burlington, USA (2010).

[3] Cancian, M.H., Rabelo, R.J: Supporting Processes for Collaborative SaaS, IFIP Advances in Information and Communication Technology, Vol. 408, 183–190 (2013).

[4] Connolly, J.: Corporate DNA: Using Organizational Memory to Improve Poor Decision Making, Journal of Management History, Vol. 16 Iss: 1, 137–138 (2010).

[5] Dahiya, N., Bhatnagar, V: Effective data warehouse for information delivery: International Journal of Networking and Virtual Organisations, Vol. 12, N° 3/2013, P 217–237 (2013).

[6] Dalkir, K.: Knowledge Management in Theory and Practice (2nd Edition). Cambridge (Massachusetts): MIT Press, London (2011).

[7] Fournier-Morel, X., Grojean, P., Plouin, G., Rognon, C. : SOA le Guide de l'Architecture du SI. Dunod, Paris (2008).

[8] Gangemi, A.: Ontology Design Patterns. Tutorial on ODP, Laboratory for Applied Ontology Institute of Cognitive Sciences and Technology CNR, Rome, Italy (2006).

[9] Gerald, K.: Cloud Computing Architecture, Corporate Research and Technologies, Munich, Germany (2010).

[10] Gronau, N.: Enterprise Resource Planning. oldenbourg.verlag, 2nd Edition (2010).

[11] Kassel, G. : Intégration de l'ontologie de haut niveau DOLCE dans la méthode OntoSpec, http://hal.ccsd.cnrs.fr/ccsd-00012203.

[12] Masolo, C., Borgo, S., Gangemi: The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. Technical Report. WonderWeb Deliverable D18 (2003).

[13] Mhiri, M., Gargouri, F. : Méthodologie de construction des ontologies pour la résolution de conflits des SI. Technique et Science Informatiques. Vol 28 (2010).

[14] Schreiber, G., Akkermans, H.: Knowledge Engineering and Management: The CommonKADS Methodology. MIT, Cambridge, USA (2000).

[15] Skaf K.M.: Application of lean techniques for the service industry. Technical report. A case study M.S., Southern Ilinois University at Cabondale (2007).

[16] Weske, M.: Business Process Management: Concepts, Languages, Architectures, Leipzig, Germany: Springer-Verlag, Berlin Heidelberg (2007).

[17] Yu, E., Horkoff, J.: Business Modeling for Business Intelligence on Data Management In Perspectives on Business Intelligence, Vol. 5, N° 1/2013, 19–32 (2013).

[18] Zahaf, S., Gargouri, F. : Mémoire organisationnelle appuyée par un cadre ontologique pour l'exploitation des processus d'affaires: 31ème Congrès en INFormatique des ORganisations et Systèmes d'Information et de Décision INFOSID, Paris (2013).

Agent Based Emergency Response Cognition Model

SAFIYE SENCER, ORHAN TORKUL, KUTLU EREN
SAKARYA UNIVERSITY

safiyesencer@yahoo.com, torkul@sakarya.edu.tr, kutlueren26@gmail.com

Abstract **-** This study considers agent based emergency response cognition model. To operate distributed and dynamic environments are inevitable in real life to realize process in a short time. Decision-making processes among autonomous agents can support to solve dynamic and large system problems. This paper presents an agent based emergency response coordination model that considers the knowledge structure, the space event type and time dimension and the dynamics of the real environment.

Keywords: Cognitive Agent Multi agent system, Emergency Response Model Cooperative Distributed Decision-Making,

I. Introduction

The cognitive cycle of the agent model try to quick situational responsiveness develop shared understanding of the operational environment, to check the condition and the execution of the strategies, to deal with multiple, simultaneous crises. The cognitive model attempts to more agile response during the making decision. The probabilistic information and response mechanism is so important part in the suggested model. The cognitive model includes the uncertainty situations for the next decision mechanism. At the same time, learning is very important in cognitive model for thinking and problem solving. Also the learning can realize the process from the experiments; information or with the methods. Moreover, the learning method able to analysis the uncertain situations with interactive markov chains in this study.

Emergency response model covers the nature events and the non-nature events. In particular, unexpected events cause the turmoil among the people most of time. To prevent the turmoil and to save their life in every time, in particular depend on the emergency responses. Appropriate responses are needed in the form of allocating resources to handle the effects of emergency responses. The form and remedy kind may help to people easily kind of the unexpected event. Also the early caution systems can able to stimulate people to this kind of events.

Determination of the unexpected events type is so important for help the people. In the World, people every time faces to nature events such earthquake, forest fires, terrorist attacks, war threads. Also to operate distributed and dynamic unexpected environments are inevitable and so important in real life to realize process in a short time. Using of the cognition model help to understand decision situations and to analyze complex cause-effect representations and to support communication. Multi-agent emergency response system has been extensively used in the different tasks of decentralized emergency response problem solving such as communication among agents, collective decision making, cooperation, collaborative planning in large scale that deals with uncertainty and conflicting information during emergency response management.

This paper organized as follows: the next section presents an overview of works in the literature related to emergency response structure, cognitive model and multi agent systems dynamic modeling and simulation; the next section introduces overall architecture which attains the proposed aims and emphasizes the roles played by the agent based emergency system components; and describes the complete working flow and details theoretical approach on which relies the work; then, the next describes how the process model applied to emergency response real model. Conclusions and future works close the paper.

II. LITERATURE SURVEY

Cognitive model compromises four components which are conative, cognitive, skills and instinctive. It covers the conative with feelings, emotions, and cognitive with thinking and problem solving, and skills with actions and driving and instinctive with reflex and reactions (Lawson, 2001). Suggested model takes in the thinking and problem solving instantaneous. In particular, learning process from experiments,

information and designed method is available with multi-agent system (Brunswik, 1957). The probabilistic information and response mechanism is so important part in the suggested model. The cognitive model includes the uncertainty situations for the next decision mechanism. The learning is very important in cognitive model for thinking and problem solving. Also the learning can realize the process from the experiments; information or with the methods. The learning method able to analysis the uncertain situations with interactive markov chains in this study.

Emergency management is so important for the sustainable qualify life conditions. In particular, to take precautions for unexpected emergency situations are so important for the human life. MAS may use in emergency situations resulting from natural and human made emergency responses, such as flood, tsunami, earthquake, terrorist attack, fire in building etc, represent complex and dynamic environments with high level of uncertainty; hence autonomous notification and situation reporting for emergency response management system will be done by multi-agent response system. Suggested agent based emergency response cognition model provides the basic components for the system. In literature, Wang et al. (2013) suggested the emergency management response system structure for a city in China. Also, Basak (2011) et al., suggested the agent based disaster management and Chou et al. 2008, suggested the dynamic parking structure with agent based platform. The suggested agent based model designed with cognitive structure.

Belief Desire Intelligence (BDI) is very important architecture for offer an agent with artificial cognitive capabilities. The other say, cognitive agent aims to perform cognitive activities with considering of the human cognitive behavior. Cognitive activities based on the three components which are perception, reasoning and execution. In addition, traditional agent characteristics which are autonomy, social ability, reactivity and pro-activities have to combine with cognitive abilities.
Some of the emergency response response systems have been developed based on multi-agents systems approach (such as: DrillSim developed by Balasubramanian et al. 2006, DEFACTO designed by Marecki et al. 2005, ALADDIN modeled by Adams et al. 2008 and Jennings et al. , RoboCup Rescue suggested by Kleiner et al. 2005, and FireGrid proposed by Berry and Usmani in 2005) and more are being developed. Our study covers the behavioral response structure which is influenced by objective (cognition or abilities) and subjective (feelings or reflexes) processes. It focuses on objective processes which may influence the behavioral response. In particular, to cognition and the relation to the decision-making process in the context of evacuation dynamics are given.

The autonomy is an ability of agent to achieve its goals without any supporting from other agents. On the other hand, the interaction of agents to get the global goal of the system is the social ability of agents. The reactivity, which is based on the relation between perception and action, is an ability of agents to respond to the environmental changes. The pro-activeness of agents is an ability to express the goal-directed behaviors. The reactions of agents to the environmental changes are the reactivity or pro-activeness that depends on what kind of architecture of agent is used to develop agents. The different characteristic of the cognitive agent in comparison with the traditional agent is the intelligence of the cognitive agent, which is shown at the improvement of the pro-activeness characteristic. Intelligence is the ability of the agent using its knowledge and reasoning mechanisms to make a suitable decision with respect to the environmental changes.

III. AGENT BASED EMERGENCY COGNITION MODEL

This section includes overall architecture which achieves the proposed aims and emphasizes the roles played by all system components; and describes the complete working flow and details. Theoretical approach on which relies the work. The cognitive cycle of the agent model presents the detail structure in agent model with algorithm. The next subtitle of the Agent Based Model gives the detail of the emergency responsive model with components.

3.1 The Cognitive Cycle of the Agent Model

The cognitive cycle of the agent model aims to quick situational awareness develop shared understanding of the operational environment, to monitor the situation and the execution of the strategies, to deal with multiple, simultaneous crises. The cognitive model aims to more agile response during the making decision.

The suggested model is able to realize communication and collaboration with coordination, coalition and collaborative information distribution properties. Agents communicate in order to achieve better the tasks of them or of the society/system in which they exist. Communication can enable the agents to coordinate their actions and behavior, resulting in systems that are more balanced. For communication, agents are able to coordinate, coalition and collaborative information each other agents.

The suggested model is able to process the data cognition of knowledge judgment system. Also, it can realize the understanding and sense making. The cognition model notations are given in Table 1.

$$a(t, x, z_n) = F^{Act}(x, z_n, K_n, D_T(t, x, z_n), Ac_T(t, x, z_n), S_T(t, x, z_n), E_T(t, x, z_n), L_T(t, x, z_n), O_T(t, x, z_n) \quad Eq(1)$$

Dynamic modeling of the abstract model for agent based emergency cognition model shows in Figure 1. Dynamic abstract model covers the action, sensing, dynamic cognition evaluation, learning and decision. In addition the control module activities are listed in Table 2. Also Eq. 1 represents the all of the activities of abstract system.

Our suggested model could be done flowing properties:

- situation assessment
- understanding context surrounding
- communication
- collaboration

Table 1. Summary of the notation

Notation	Description
a	Agent's dynamic abstract model
A= {x, y, $z_{n,n,n}$}	Set of all agents
t	time
x	bid, information
Z_n	related system activity {sending bid; receiving bid; negotiate; notify; request notification; computing; learning
F^{Act}	Agent's all of the functions
Kn	collaboration
D	process (task)
Ac	action
S	sensing of the situation
E	evaluate (alternatives)
L	learning (situation)
O	Outcomes of task T
B	Update (B,s)
T	Compare(B,D,E)
An	Announce
Ne	Negotiate
N	Notify
R	Request notification
C	Computing
Co	Coalition
Coo	Coordination
CIS	Cooperative Information Sharing
Col	Collective Information

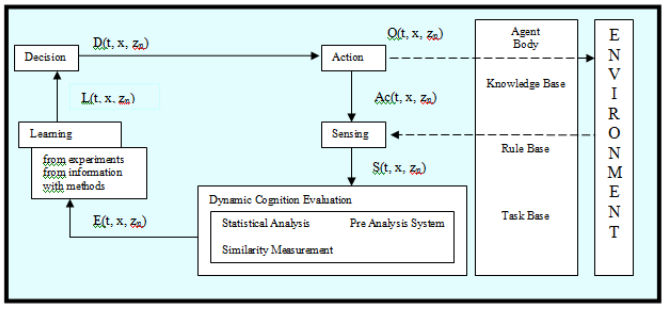


Figure1. Dynamic modeling of the abstract system

III.2 Emergency Response Agent Based Cognition Model

In this paper, the current emergency response management and response systems are analyzed and taking in consideration of domain requirements, agent-based design methodology and systems' comparative view. Dynamic probabilistic unexpected movements and service capabilities illustrated with dynamic stochastic model. Modeled of the system includes the expected and unexpected events. Markov chain model is applied to system and solved for the steady state probability distributions of queue lengths and waiting times. Several performance measures computed with the developed algorithms for the modeled system.

Suggested agent model covers the some parts, which are collective information, coordination, communication, coalition, cooperative distributed information sharing, problem solving and evaluation, decision making. Contract Net (CNET) is preferred to achieving efficient cooperation through task sharing in networks (Weiss, 2012) as well as used in Emergency response System for communicating problem solvers.

Multi-agent emergency response system includes the different tasks of distributed emergency response problem solving such as collective decision making, communication among agents, cooperation, collaborative planning in large scale that deals with uncertainty and conflicting information during emergency response management (Adams et al.). In detail, the suggested type of emergency response systems can be viewed on information and knowledge fusion and take the feedback from the existing agents for sensing, coordinating, decision making and acting. It must be able to achieve these objectives in environments in which: control is distributed; uncertainty, ambiguity, imprecision; multiple agents with different aims

and objectives are present; and resources are limited and vary during the system's operation.

The suggested model may provide some benefits which are); more robust, interoperable, and priority sensitive communications, better situational awareness, improved decision support and resource tracking, greater organizational agility, better engagement of the public. Emergency response model covers the very wide area respects of the nature events (earthquake, floats, fires, hurricane, and thunder storms), terror events (considering of the coming information), fault of the people or devices (accidents, fires, explosions etc.).

Knowledge Base Agent: It includes the terror events and accidents as unnatural events, also natural events include the earthquake, flood, fire, hurricane, statistical data and current data. The knowledge base agent takes data from environment observation. Also system inputs connected with the knowledge base with simultaneously coming data information, historic data and weather data. At the same time, learning process shares the data with system inputs and knowledge base. It covers the database, which is related to the system components. It provides the using of the information whole system.

Input Agent: In this part uses the knowledgebase agent's database and obtained simultaneous data. At the same time statistical data and simultaneous data are evaluated with together in this part.

Risk Assessment Agent: It occurs from two parts, these are analysis and evaluation. Analysis part considers the hazard analysis, damage assessment, loss assessment with current situation in grading to the importance degree during the decision making. Evaluation includes the alternatives and criteria's comparison structure.

Learning Agent: Learning agent is the most important part. Three different approach help to realizes the learning procedure which are experiment, information and method also it can assist to capability of the intelligence in agent system.

Performance Estimation Agent: It includes the information and coming data information. Processing of the information provides the coming data analysis and evaluation with proper decision making structure.

Service Provider Agent: The agent includes the service system which includes the ambulance service, logistic service, police linked service, combined emergency service.

Decision Making Agent: It includes the system outputs which are prediction of the event type, prediction of the location, prediction of the correct resource usage and correct resource coordination.

The system goal in this agent-based coordination network between emergency support system and people who are needs to support is divided into three parts: defining the initial negotiation strategy, considering the emergency assessment factors, and finding the cost functions and report results.

The architecture of a cognitive agent shown in Fig. 2 consists of five modules: perception, decision making, knowledge, control, and communication. The perception module is responsible for data acquisition in the environment. The decision making module is in charge of the agent making a decision in an autonomous way.

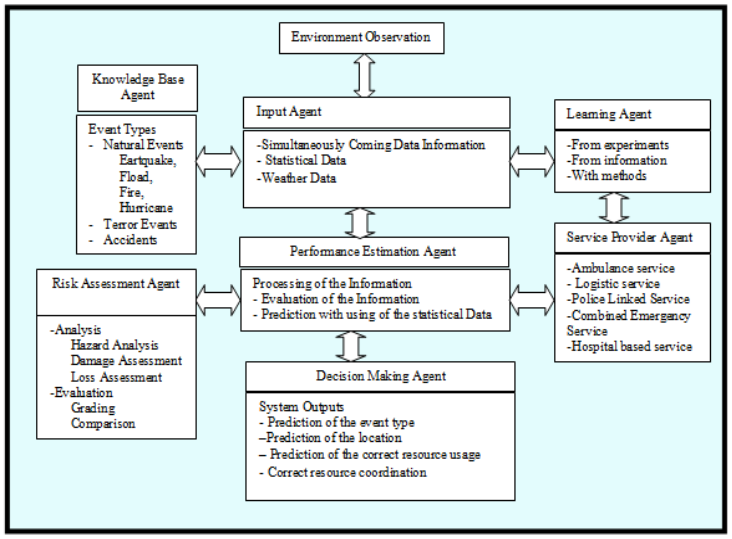


Figure 2 Agent Based Emergency Response Cognition Model

The control module processes the plan into tasks and executes the tasks to the environment. At the same time, this module sends the tasks to the communication module if the plan is processed by the agent team. The communication module is responsible for interactions between the cognitive agent and the agent community. It receives or sends messages, interprets them and transmits the tasks of the control module to other agents. The knowledge module contains

intentions, and plans of the agent. Control module is so important for the decision making agent. It covers the update, process of the task, evaluate of the alternatives, compare, learning, announce, negotiate, notify, request, notification and computing. Figure 3 represents the common model for markov chain based probabilistic transition and information. The algorithm structure shows the main loops of the control module. Some of the main processes in decision making algorithm are listed in below:

Process D(t,x,z_n)
D:process(task)

Sensing SD(t,x,z_n)

Evaluate E(t,x,z_n)
e:evaluate(alternatives)
Evaluate
Ranking of the process results

Learning L(t,x,z_n)
l:learning(situation)
Obtain rules from data with clustering, classification and ranking

Announce Situation assessment
Evaluate
Compare

Negotiate Understanding the context surrounding
Compare
Evaluate
Learning

Notify Communication
Announce
Negotiate
Notify
Sharing

Compare t:=compare

Computing c:=computing(alternatives)

Filter f:=filter (B,D,I); /* agent select goals*)
Filtering of the current situation
Connection of the filtering situations and alternative plans

Plan p:=p
Alternative plans

Suggestion of the new plan for current situation
Realization of the plan degree
Comparing of the plan and the real situation
Plan's success degree

Execute (p) adjusting of the conditions and situations for execution of the plan

```
B:= B_initial; /*initial values of beliefs*/
  Input probabilistic transition system (S,→)
        Reward function ℜ:S→ℝ
      Output S/~pℜ
  Method Bid:=∪vεℜ+{(p'εS|-ℜ(p')=v}|-{∅};
      Spl:=Bid;
      While Spl not empty do
            Choose C in Spl;
            Old:=Bid;
            Bid:=P-Refine(Bid,C)
            New:=Bid-Old;
            Spl:=(Spl-{C})∪New
       Odd
       Return Bid;
    B:= B_initial; /*initial values of beliefs*/
    I:= I_initial; /*initial values of goals*/
Loop
   a:=see; /*agent gets information from sensors*/
   n:=notify /* agents take information
   if a:=disturbance (type) then
  send(message); /* the emergency system cannot be operated, and the disturbance
  belongs to  categories either long recovering time or unable recovery, so the
  agent requires the MES rescheduling*/
      Else
          B:= update(B, a); /*agent updates its beliefs*/
          D:= process(task); /*agent gets all possible states of the task*/
          C:=computing (alternatives); /* agent evaluate the alternatives*/
           t:= compare(B,D); /*agent compares between beliefs and desires*/
           l:= learning (b,s);
           If t:=0 (normal status) then
           send(message); /*agent reports the normal status of the environment*/
           else
           I:=filter(B, D, I); /*agent selects goals*/
           p:= plan(B,I); /*agent has a plan to get goals*/
           a:= evaluate(alternatives0
                If not empty (p) then
           p:=plan(B,I); /*agent has a plan to get goals*/
             If not empty(p) then
             execute(p);/*agent executes the plan*/
             l:= learning(situation)
              send (message); /*agent sends a message to the related agents for
              updating the new plan*/
                 updating the new selection*/
             else
             cooperate(agents);/*agent cooperates with the other agents*/
             a:=select(agent); /*agent selects another agent to do its work*/
                 If not empty(a) then
                 Coordination (task)
                 send(task); /*agent sends its work to the selected agent*/
                 else
              send(message); /*agent sends a message to the require
             rescheduling*/
              End if;
          End If;
          End If;
      End If;
End Loop.
```

Figure 3: A common model for information processing. Example schema of information processing unit.

IV. AGENT BASED EMERGENCY RESPONSE COGNITION MODEL: CASE STUDY

The case study covers the some steps which are system analysis and design; building of conversation; system processes.

4.1 System analysis and design

The coordination network covers the system component, experts and people. FIPA –Contract-Net-Manager protocol is applied to analyze and

design the system. Finally, the pattern of this system is built and illustrated. This study emphasizes the coordination of the network emergency assist system and who needs the assist one. The agent mechanism structure can be applied to establish such a negotiation system with the distinction from the emergency information system, and at the same time based on research as shown in the literature review.

4.2 System analysis

The agent can complete complicated negotiation with related agent. Agents work automatically without rest on coordination, enabling the drivers to seek out. As the agent architecture, the FIPA – Contract-Net-Manager protocol is preferred.

.

4.2.1 Agent classes

In negotiation, the agents engage in dialogue, exchanging proposals with each other, evaluating other agents' proposals and modifying their own proposals and then modifying their own proposals until all agents are satisfied with the set of proposals. Standard negotiation mechanisms adopted are based on game theory or on human-inspired negotiations. Every task also defines higher level, complex interaction protocols requiring coordination between multiple agents. First, the system checks emergency assist opportunities information for user when he inputs data into the system, and sends a message to emergency support agent. Second the user agent negotiates with the user agent through to emergency assist agent. The bidding continues until the driver agent accepts one bid or rejects them all.

4.2.2 Tasks

A task is a structured set of communications and actions. The ovals denote tasks that the role must execute in order to accomplish its goal. These concurrent tasks are defined as a finite state automation specifying messages sent between roles and tasks. The lines between nodes indicate protocols between tasks, which define a series of messages between the tasks that allow them to work cooperatively.

In this step, an agent class diagram is created, as depicted in Figure 4 from the viewpoint of the roles, documented. The agent class diagram depicts agent classes as boxes, and the conversations among them as lines connecting the agent classes. Four agent classes are defined: user agent; assist agent; information agent; decision making agent.

4.2.3 Roles

- Reporting any situation that requires a police officer at the scene (e.g. assaults, traffic accident, burglary report, damage to property, parking complaint, other ordinance violations, etc.)
- Call an ambulance for medical assistance.
- Reporting fire, smoke or fire alarm.
- Reporting a crime in progress.
- Reporting suspicious or criminal activity. (shouts for help, glass breaking, vehicle or person that does not appear to belong in neighborhood).

4.2.4 Sequence Diagrams

A use case is a narrative description of a sequence of events defining desired system behavior. A sequence diagram depicts a sequence of events between multiple roles and defines the minimum communication between roles. Use cases can be drawn from the system requirements and users. Then the use cases can be restructured into sequence diagrams. The proposed system has five main sequences of events, which are described in the following by use cases and sequence diagrams.

Refining roles: The third step is to ensure that all the necessary roles have been identified, and to develop the tasks that define role behavior and communication patterns. Through applying the use case step, the roles of the proposed system have been defined roughly, so in this step they are refined, and tasks associated with each role are created. Fig.4 illustrates a agent architecture role model.

4.2.5 Conversations

With negotiation, the agents engage in dialogue, exchanging proposals with each other, evaluating other agents' proposals and then modifying their own proposals until all agents are satisfied with the set of proposals. Standard negotiation mechanisms adopted are based on markov chain approach and human inspired negotiations.

In negotiating phase, the roles which are emergency support agent, decision making agent, user agent includes the main messaging system. A communication diagram is a pair of finite state machines defining a conversation between two participant agent classes. The

syntax of the communication class diagram is very similar to that of the roles include emergency response parameters.

The assist agent realizes the negotiation as follows:
The initiator agent estimates the minimum cost and time resource point
The respondent agent proposes the bid lately
The initiator agent proposes the minimum cost
The initiator agent proposes the minimum distance
The initiator agent bids

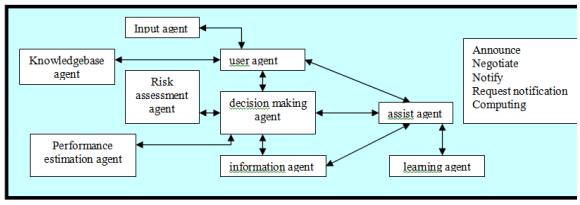


Figure 4 Agent based architecture

The suggested model realizes the following steps:
Step 1: Determine the asking help choices and the assist alternatives
Step 2: Revise asking resource slightly
Step 3: Diagnose whether exceeding the end time and distance
Step 4: Consider the types of messages
Step 5: Control of the limitations
Step 6: Evaluate the final situation

The agent architecture defines the configuration of the system to be implemented. The overall system architecture is defined by deployment diagrams. The proposed system is divided into two subsystems. Subsystem 1 includes four agents: agent; assist agent; information agent; decision making agent that provides the negotiable spaces of the dynamic behavior. Also some of the system components uses the some input data such as calling system people's voice tone like calm, angry, excited, slow, rapid, soft, loud, vulgar, laughing, crying, normal, distinct, slurred, intoxicated, nasal, stutter, lisp, raspy, ragged, clearing throat, deep breathing, cracking voice, disguised, accent, electrically altered, familiar, rational, irrational; background voice frequency like, airport, animal noises, baby, clear, local, school, factory machinery, office machinery, restaurant, television, house noises, motor, music, street noises, kids, traffic, long distance, party

4.2.6 System process

In multi agent decision making, the agents utilize classical artificial intelligence decision making methods to decision making their activities and resolve any foreseen conflicts.

Alternatives	Event Type	Event Distance	Emergency Size	How many pec

Table 4 Emergency response criteria types

In order to represent the subjective multiple attribute preference of making decision of emergency, a decision hierarchy based on the ordered weighted averaging (OWA) is used. OWA is a useful decision analysis tool which analyzes the decision problem quantitatively by utility. The basic idea of the OWA is for dealing with a problem where its result comes from two or more attributes. Also OWA provides a general class of parameterized aggregation operators that include the min, max, average, and several other operators. OWA was originally introduced by Yager (1988; 1993) has gained much interest among researchers, hence many applications in the areas of decision making, expert systems, data mining, approximate reasoning, fuzzy system and control have been proposed.

Attributes can be specified and then a ordered weight function can be constructed to evaluate utility values of different solutions. The solution with the best utility will be selected. Each agent keeps a driver's decision hierarchy as multi-attribute preferences. The people who have some problems that related about the

1. System have to determine the problem class
2. After classification, system has to find solution about the problem in a short time and short distance to someone who needs to emergency assist.
3. Emergency assist agent system has to making decision with the all of the com-ponents.

The suggested system considers the different parameters for the evaluation of the information for selection of the best choice. The system decision making process uses the different management methods like the OWA and fuzzy logic for the uncertainty situations. Also multi attribute utility theory uses the certain information and uncertain information with fuzzy approach. $U=w_1U_1+\ldots+w_iU_{i+\ldots+}w_nU_n$, where U is the overall utility value; w_i is the

weight of the i th attribute; and U_i is the utility of set i th attribute. The sum of the weights $w_1+\ldots+w_n$ is equal to one. Weights correspond to the relative importance that manager place on the related attributes.

V. Conclusion

This paper discusses agent based emergency response cognition model with an active multi agent database system which incorporates active rules in a multi computing environment. The partitioning of the rule set into multi agent system events has also been obtained from Markov chain model. Answer based event recognition has been introduced to active multi agent databases, which is an important contribution from the perspective of performance. (matching the situations) This system helps people to reach emergency remedy resources easily. Finally, due to frequent changes in the positions and status of objects in an active mobile database environment, the issue of temporality should be considered by adapting the research results of temporal database systems area into active mobile databases.

This paper gets the agent-based simulation and determines an optimal plan to emergency response model in the shortest time possible. Agent simulation for emergency response cognition model improves upon other simulation models that are concerned with numerical analyses of inputs or amounts of people and structures. The agent-based system for emergency response model is grounded on empirical data taken from real-world experiments. If the agent sees an exit, it will proceed towards it and if it receives any types of direction to leave, that will be carried out without failure. Further study includes the improvement of the text mining techniques with new respect. Also agent based emergency response cognition model provides to evaluate uncertain and vagueness information.

References Adams M., N.M., Field, E. Gelenbe, D.J. Hand, N.R. Jennings, D.S. Leslie, D. Nicholson, S.D. Ramchurn, S.J. Roberts, A. Rogers, "The ALADDIN Project: Intelligent Agents for Emergency response Management".

Adams, M. Field, E. Gelenbe, D. J. Hand, "The Aladdin Project: Intelligent Agents for Emergency response Management - IARP/ EURON Workshop on Robotics for Risky Interventions and Environmental Surveillance", 2008.

Balasubramanian , Massaguer, Mehrotra , Venkatasubramanian, "DrillSim: A Simulation Framework for Emergency Response Drills"; Proc. of ISCRAM, 2006.

Basak, S., Modanwal, N., Mazumdar, B.D., Multi-Agent Based Disaster Management System: A Review, IJCST Vol. 2, Iss ue 2, June 2011.

Berry, D., Usmani, "A. FireGrid: Integrated Emergency Response and Fire Safety Engineering for the Future Built Environment"; UK e-Science Programme All Hands Meeting, Nottingham, UK, Sept. 19-22, 2005.

Brunswik, E.. Scope and Aspects of Cognitive Problems. Contemporary Approaches to Cognition, pages 5–31, 1957.

Chou, S.Y., Lin, S.W., Li, C.C., Dynamic parking negotiation and guidance using an agent-based platform, Expert Systems with Applications, Vol.35, Is. 3, 2008, pp 805-817.

FIPA, ''Specification Part 2:Agent Communication Language,'' The text refers to the specification dated 23 October 1997.

Jennings, N., Gopal Ramchurn, Mair Allen-Williams, Raj Dash, Partha Dutta, Alex Rogers, Ioannis Vetsikas; "The ALADDIN Project: Agent Technology to the Rescue".

Khaled M. Khalil, M. Abdel-Aziz, Taymour T. Nazmy, Abdel-Badeeh M. Salem, "Multi-Agent Crisis Response systems – Design Requirements and Analysis of Current Systems".

Kleiner, B. Steder, C. Dornhege, D. Höfer, "RoboCupRescue - Robot League Team RescueRobots Freiburg (Germany)"; RoboCup (Osaka, Japan), 2005.

Lawson, B. The Language of Space. Architectural Press (Space and Time), 2001.

Marecki, N. S., Tambe, M., "Agent-based Simulations for Emergency response Rescue Using the DEFACTO Coordination System"; Emergent Information Technologies and Enabling Policies for Counter Terrorism, 2005.

RoboCup project, [Online] Available: http://www.robocup. org/

Wang, D.Y., Pan, L.W., Lu, L., Zhu, J.P., Liao, G. X., Emergency Management Business Process Reengineering and Integrated Emergency Response System Structure Design for a City in China, Procedia Engineering 52 (2013) 371 – 376

Weiss, G. "Multi-Agent Systems – A Modern Approach to Distributed Artificial Intelligence", Reprint: 2001, Massachusetts Institute of Technology; pp. 83, 87, 88, 125, 127, 129, 242, 356, 360.

Yager, R.R., 1988. On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Transactions on Systems, Man and Cybernetics, B 18: 183-190.

Yager, R.R., 1993. Families of OWA operators, Fuzzy Sets and Systems, 59: 125-148.

Multi Hybrid Keyword Processing for Topic Decision of Unstructured Data

Jinwoo Lee, Hyoungmin Ma, Gitae Lee, Kihong Ahn, Sukyoung Kim

***Abstract*— Amount of information and difficulty of the user's information selection has direct proportion relation. Also title is consists of exaggerated expression. Since, authors want to summarize about document. Therefore title is almost different from contents. If these case are more increased, offering information by simple keyword search will be reached to the limit. In this study, to solve these problems, we applied TF-IDF to extract keyword in particular documents which have scarcity words in all documents and applied LDA algorithm for to find topic about single document. Finally, we have proposed the methodologies that add description on scarcity word and topic through to extract the Trigram of the entire document. In this study, to verify the accurate of methodology, we made supervised data and compared these data with data that made by suggested methodology.**

I. INTRODUCTION

Development of WEB 2.0 environment increase diversification and complexification form and expression of the information which is made in sudden expansion of SNS. This is the important reason to the users why fail to find accurate information. Particularly, the redundancy of signification and the metaphor expression to be elements which obstruct the satisfaction of searching information. Expansion of cloud technology is possible to make videos, photos and document information of pdf file infinitely and easily. So it make more difficult to find information. Also the sentences in SNS (like twitter, facebook) are abbreviated form and unphotographic information. So it is generate the limit to find information with few keywords. These problems make the time consuming for searching information. In order to our study solves this problems which are mentioned sentence before, we apply TFIDF to extract keyword for scarcity word in documents and apply LDA algorithm to find a TOPIC in single document. Finally, we suggest additional explanation methodology of scarcity word and Topic to extract Trigram in all documents and compare result of experiment to extract Trigram for verifying accuracy of our methodology.

Lee JinWoo is with Department of Computer Engineering, University of National Hanbat, Daejoen, South Korea (e-mail: fniko0084@gmail.com).

Ma HyoungMin is with Department of Computer Engineering, University of National Hanbat, Daejoen, South Korea (e-mail: mahm0000 @naver.com).

Lee GiTae is with Department of Computer Engineering, University of National Hanbat, Daejoen, South Korea (e-mail:mm1023@naver.com).

Ahn KiHong is with Department of Computer Engineering, University of National Hanbat, Daejoen, South Korea (e-mail: khahn@hanbat.ac.kr).

Kim SuKyoung is with Department of Computer Engineering, University of National Hanbat, Daejoen, South Korea (e-mail: kimsk@hanbat.ac.kr).

II. RELATE RESEARCH

A. TF-IDF

TF–IDF is the product of two statistics, word frequency and inverse document frequency. There are various ways for the extracting two values (TF, IDF). In the case of the word frequency tf(t,d), the simplest selection is to use the raw frequency of a word in a document, i.e. the number of times that word t occurs in document d. If we denote the raw frequency of t by f(t,d), then the simple tf scheme is tf(t,d) = f(t,d)

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (1)$$

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max(f(w,d) : w \in d} \quad (2)$$

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

The inverse document frequency is a measure of whether the word is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the word, and then taking the logarithm of that quotient. A high weight in TF-IDF is reached by a high word frequency (in the given document) and a low document frequency of the word in the whole collection of documents; the weights hence tend to filter out common words. Since the ratio inside the idf's log function is always greater than or equal to 1, the value of idf (and TF-IDF) is greater than or equal to 0. As a word appears in more documents, the ratio inside the logarithm approaches 1, bringing the idf and tf-idf closer to 0.

B. LDA(Latent Dirichlet Allocation)

When there exist the parameter of any probability distribution, LDA is Generative Model of the viewpoint that generate data based on random process. If we know topic distribution of document and each words to generate probability, we can calculate specific document probability.

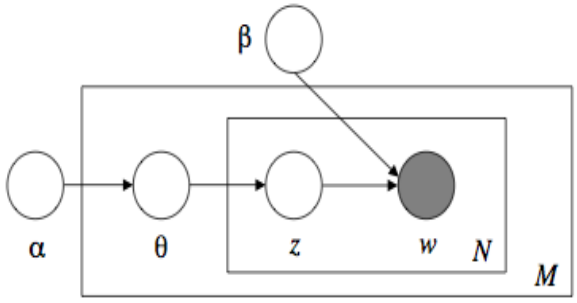


Fig. 1. LDA's concept diagram.

Latent Dirichlet Allocation given number of the M documents, it based that the documents has few existing k topic. At first, to use probability distribution at the model is as follows.

ϕ: The word distribution for topic k

z_{ij}:The topic for the jth word in document i(index)

w_{ij}:The jth word in document i(index)

In here, w_{ij} is given through the actual document, other potential variables can't observed. It is a potential variable which other variables can not be observed.

$\theta_i \sim Dir(\alpha)$:Follow the k dimension Dirichlet distribution.

$z_{ij} \sim Mltinomial(\theta)$: Follow the multinomial distribution.

w_{ij} follow generated word probability by topic that pointed by z_{ij} . At that time α is Dirichlet distribution and β is $k \times V$ matrix parameter that contain word generate probability. About topic that pointed by z_{ij} w_{ij} is conditioned by the word generating probability $p(w_{ij} \mid z_{ij}, \beta)$. At this time, α is the parameter of Dirichlet distribution and β is the probability of topic k that can give with each result V which is also calculated as $k \times V$ in the matrix. This model can be interpreted as follows. For each document, they have weight for number of k subject and z_{ij} subject of each word that chosen in multinomial distribution of weight. Finally, real words w_{ij} are selected based on specific topic.

C. N-gram

It is necessary to process of lexical for understanding sentences but the common grammar of language is very complex also many common users don't follow the standard grammar. There are various algorithms that used to analyze like these sentences. In these algorithms, n-gram has more fast and simple handling advantages than other algorithms. It is the language model which is possible to calculate the meaning whether it is real with the word link of number of n.

III. RESEARCH

A. *Basic data*

Basic data is made by normal people. So it is very similar to the data on web which we can see easy. To process these data, computer need many amount of preprocessing steps because the form is not defined. Also the Korea grammar has postposition, so the word's form changes to various form. As a result, we need to divide noun or find infinitives. It is more difficult than processing English. To solve these problem in this research extract verb and no use Korean parser (Komoran-1.12) to extract verb and noun. And also makes stop word dictionary to prove extracted noun and infinitives. The stop word dictionary is composed 1230 words including unknown meaningless and abstract word, article and postposition. Finally, processed documents are consists of only words by preprocessing. This data is processed by TF-IDF algorithm and Topic modeling algorithm which called LDA. The data was accumulated for 3 years (2011~2013). And also ideas are composed 548 ideas in 2013, 266 ideas in 2012, and 447 ideas in 2011. Each idea are composed in their background of occurrence, necessity, technical core and scenario. Therefore 1261 data are used to this research. Next figure is value of every year's exaggerated topic. In the fig 2, the mismatching ratio between contents and title is 50% in 2013, 56% in 2012, and 38% in 2011. The document of more abstractive form show the more mismatching probability. As a result, when they saw title, they can't inference about document topic.

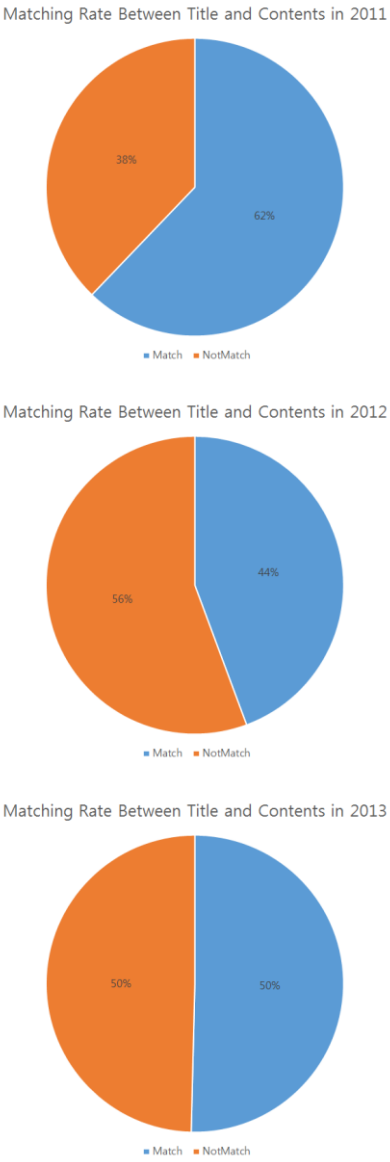


Fig. 2. A value of every year's exaggerated topic. (2011, 2012, 2013)

TABLE I
LIST OF EXTRACTED KEYWORD WHICH USING TFIDF ALGORITHM

Word	TF-IDF
컨텐츠(contents)	0.3868727611912199
헤드셋(headset)	0.1889129911547502
컴퓨터(computer)	0.0936188338596118
불면증(insomnia)	0.0910712525356702
집중력(concentration)	0.0558887009703448
우울증(melancholia)	0.0513390289607587
헤어밴드(hair bands)	0.0316357938822197
긴장감(tension)	0.0230024633755598
스트레스(stress)	0.0184689764454966
스마트(smart)	0.0154025990070782

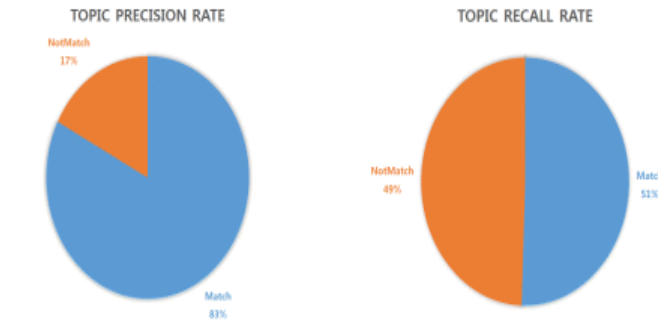


Fig. 3. Precision and recall of topic (left-precision, right-recall).

B. *Keywords extraction*

Each document contains their representative keywords. But it is very hard to find representative keywords in huge amount of word. By using TF-IDF algorithm, top 20 keywords are extracted

after preprocessing (ex. Korean parser, stop word dictionary). Table 1 is lists of extracted keywords which using TF-IDF algorithm at no.1 data.

C. Topic Modeling

In this chapter, we offer the result to find key word with TF-IDF's result to raise the key word's reliability. This method extract topic from each documents through topic modeling algorithm (LDA). Of course all documents already parsed and extracted noun and infinitives by morphological analyzer. To verify these keywords, we have extracted topic by supervised basic data. Total counts of verification documents are 430, also count of processed documents are 430. They are supervised data and LDA data. The standard of comparison is whether appear supervised topic in extracted topic by LDA. To measure this research's likelihood, each documents are processed EM-Algorithm 1000 times. As a result of algorithm, each document are normally included 5 keywords. These words be representative word in document.

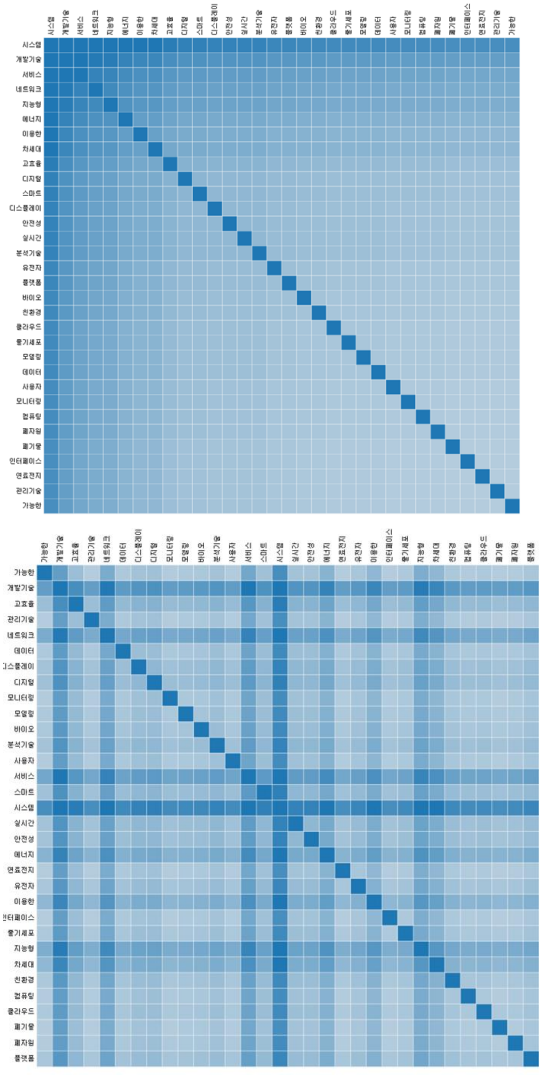


Fig. 4. Les Miserable Co-occurrence graph of relation between words (top-Sort by frequency, bottom-Sort by name)

D. Clustering

We can't know specific meaning of only a word. That is reason generate needs for analysis of sentence level. In this paper, we clustered word by trigram methodology for founding relation between words. This procedure that show relation words between tf-idf result and LDA result can solves problem for ambiguous word in context. Trigram expression is follows:

$$[PR_n Ek_n] + [Ek_n] + [PO_n Ek_n] \tag{4}$$

Clustering result of trigram can show relations between words. If frequency of trigram word has high variable, it can suppose high relation of these words. In fig 4, it shows relation of top 30 frequency words through Les Miserable Co-occurrence graph. Extracted words like table 2 are related to the middle word. In these trigram words, we select extracted topic word and directly related trigram word, and it must provide topic word as additional description word.

$$R(W \mid T) = f(w_n) + \sum_{i=1}^{K} g(\alpha_i, w_n) \tag{5}$$

(5) is equation of calculating relation between trigram and topic. w_i is one of the trigram words that maked by topic T. f(w,T) is rate of specific word w in trigram that make by topic T. i.e. f(w,T) is binomial distribution function,

$$f(w) = \frac{n(w)}{\sum_{i=1}^{n} N_i} \tag{6}$$

$$g(\alpha, w) = P(w \mid \alpha) \quad (7)$$

We select w which maximum value in R(w|T) and it is describe topic.

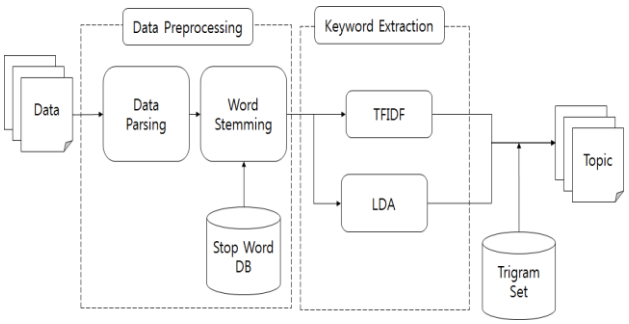


Fig. 5. Entire process model.

IV. CONCLUSION

In this paper, we show that provide keyword in document to user through TF-IDF and LDA algorithm about unstructured data. Morpheme logical analysis and word stemming through stop word dictionary improve result of our procedure. Also we make supervised data for proving unsupervised data to measure precision and recall. As a result we can improve high precision. On the other hand, recall cannot reach expected point. Extracted trigram for fool recall, we also suggested a methodology for measuring relation between topic and word. In the future, we expect reached high precision and recall as adapt this methodology.

There are several directions we plan to investigate in the future. One is making abstract word dictionary that impede recall. Another one is adapt trigram methodology for high quality. We expect to use this methodology for information select to any user that can easily select information when they want.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) 993-1022

[2] Wiliam B. Cavnar, John M. Trenkle, "N-Gram-Based Text Categorization ",Environmental Research Institute of Michigan P.O. Box 134001 Ann Arbor MI 48113-4001

[3] Juan Ramos, "Using TF-IDF to Dewordine Word Relevance in Document Queries" Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855

[4] Chenghua Lin, Yulan He, Richard Everson, "Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE TRANSCATIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO.6, JUNE 2012.

[5] Seungil Huh, Stephen E. Fienberg, "Discriminative Topic Modeling Based on Manifold Learning", ACM Transactions on Knowledge Discovery from Data, Vol. 5 No. 4, Article 20, Publication date: February 2012

[6] Aurora Pons-Porrata, Rafael Berlanga-Llavori, Jose Ruiz-Shulcloper, "Topic discovery based on text mining techniques", Information Processing and Management 43 (2007) 752-768

[7] A. P. Dempster, N. M. Laird, D. B. Rubin "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Vol.39,No.1(1977),pp.1-38