SESSION

INTELLIGENT SYSTEMS, INFORMATION EXTRACTION AND ENGINEERING, DATA PROCESSING, AND APPLICATIONS

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

A Prudent-Precedence Concurrency Control Protocol for High Data Contention Main Memory Databases

Mohammed Hamdi[1], Weidong Xiong[1], Feng Yu[2], Wen-Chi Hou[1]
[1]Department of Computer Science, Southern Illinois University, Carbondale, IL 62902
[2]Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555
(mhamdi, wxiong)@siu.edu, fyu@ysu.edu, hou@cs.siu.edu

Abstract - *In this paper, we propose a concurrency control protocol, called the Prudent-Precedence Concurrency Control (PPCC) protocol, for high data contention main memory databases. PPCC is prudently more aggressive in permitting more serializable schedules than two-phase locking. It maintains a restricted precedence among conflicting transactions and commits the transactions according to the serialization order established in the executions. A detailed simulation model has been constructed and extensive experiments have been conducted to evaluate the performance of the proposed approach. The results demonstrate that the proposed algorithm outperforms the two-phase locking in all ranges of system workload.*

Keywords: Concurrency Control, Main Memory Database, Serializability, Serialization Graph, 2PL

1 INTRODUCTION

During the past few decades, there has been much research on currency control mechanisms in databases. The two-phase locking (2PL) [7], timestamping [3, 4, 13], and optimistic algorithms [10] represent three fundamentally different approaches and have been most widely studied. Many other algorithms are developed based on these or combinations of these basic algorithms. Bernstein et al. [2] contains comprehensive discussions on various concurrency control protocols.

Optimistic concurrency controls (OCCs) have attracted a lot of attention in distributed and real time databases [8, 9, 11, 12, 5, 6] due to its simplicity and dead-lock free nature. Transactions are allowed to proceed without hindrance until at the end - the verification phase. However, as the resource and data contention intensifies, the number of restarts can increase dramatically, and OCCs may perform much worse than 2PL [1]. As for the timestamp ordering methods, they are generally more appropriate for distributed environments with short transactions, but perform poorly otherwise [14]. 2PL and its variants have emerged as the winner in the competition of concurrency control in the conventional databases [1, 5] and have been implemented in all commercial databases.

Recent advances in wireless communication and cloud computing technology have made accesses to databases much easier and more convenient. Conventional concurrency control protocols face a stern challenge of increased data contentions, resulted from greater numbers of concurrent transactions. Although two-phase locking (2PL) [7] has been very effective in conventional applications, its conservativeness in handling conflicts can result in unnecessary blocks and aborts, and deter the transactions in high data-contention environment.

Most database management systems now are designed assuming that data would reside on disk. However, as hardware technology advances, memory capacity quickly increase while the price drops dramatically. The computers can nowadays easily accommodate tens or even hundreds of gigabytes of memory with which to perform operations. A significant amount of data can be held in the main memory database without imposing severe constraints on memory size. Thus, the use of main memory databases is becoming an increasingly more realistic option for many applications. Database researchers can exploit main memory database features to improve real-time concurrency control because the database system optimized for main memory can provide speedy real-time concurrency control outcomes and support much higher transaction throughputs [15].

In this paper, we propose a concurrency control protocol, called prudent-precedence concurrency control (PPCC), for high data contention main memory databases. The idea comes from the observations that some conflicting transactions need not be blocked and may still be able to complete serializably. This observation leads to a design that permits higher concurrency levels than the 2PL. In this research, we design a protocol that is prudently more aggressive than 2PL, permitting some conflicting operations to proceed without blocking. We prove the correctness of the proposed protocol and perform simulations to examine its performance. The simulation results verify that the new protocol performs better than the 2PL and OCC at high data contention environments. This method is also simple and easy to implement.

The rest of this paper is organized as follows. In Section 2, we introduce the prudent-precedence concurrency control protocol. In Section 3, we report on the performance of our protocol. Conclusions are presented in Section 4.

2. THE PRUDENT-PRECEDENCE CONCURRENCY CONTROL

To avoid rollback and cascading rollback, hereafter we assume all protocols are strict protocols, that is, all writes are performed in the private workspaces and will not be written to the database until the transactions have committed.

2.1 Observations

Our idea comes from the observation that some conflicting operations need not be blocked and they may still be able to complete serializably. Therefore, we attempt to be prudently more aggressive than 2PL to see if the rationalized aggressiveness can pay off. In the following, we illustrate the observations by examples.

Example 1. Read-after-Write (RAW). The first few operations of transactions T_1 and T_2 are described as follows:

T_1: $R_1(b)$ $W_1(a)$ ..., T_2: $R_2(a)$ $W_2(e)$ ...,

where $R_i(x)$ denotes that transaction i reads item x, and $W_j(y)$ denotes that transaction j writes item y. Consider the following schedule:

$R_1(b)$ $W_1(a)$ $R_2(a)$ ...

There is a read-after-write (RAW) conflict on data item "a" because transaction T_2 tries to read "a" (i.e., $R_2(a)$) after T_1 writes "a" (i.e., $W_1(a)$). In 2PL, T_2 will be blocked until T_1 commits or aborts. T_2 can also be killed if it is blocked for too long, as it may have involved in a deadlocked situation.

If we are a little more aggressive and allow T_2 to read "a", T_2 will read the old value of "a", not the new value of "a" written by T_1 (i.e., $W_1(a)$), due to the strict protocol. Consequently, a read-after-write conflict, if not blocked, yields a precedence, that is, T_2 precedes T_1, denoted as T_2 -> T_1. We attempt to record the precedence to let the conflicting operations proceed.

Example 2. Write-after-Read (WAR). Consider the same transactions with a different schedule as follows.

$R_1(b)$ $R_2(a)$ $W_1(a)$ ...

Similarly, $W_1(a)$ can be allowed to proceed when it tries to write "a" after T_2 has read "a" ($R_2(a)$). If so, the write-after-read (WAR) conflict on item "a" produces a precedence T_2 -> T_1 in the strict protocol. Note that T_2 again reads "a" before T_1's $W_1(a)$ becomes effective later in the database.

Precedence between two transactions is established when there is a read-after-write or write-after-read conflict. Note that a write-after-write conflict does not impose precedence between the transactions unless that the item is also read by one of the transactions, in which case precedence will be established through the read-after-write or the write-after-read conflicts.

Note that either in a read-after-write or write-after read conflict, the transaction reads the item always precedes the transaction that writes that item due to the strict protocol.

2.2 Prudent Precedence

To allow reads to precede writes (in RAW) and writes to be preceded by reads (in WAR) without any control can yield a complex precedence graph. Detecting cycles in a complex precedence graph to avoid possible non-serializability can be quite time consuming and defeat the purpose of the potentially added serializability. Here, we present a rule, called the Prudent Precedence Rule, to simplify the graph so that the resulting graph has no cycles and thus automatically guarantees serializability.

Let $G(V, E)$ be the precedence graph for a set of concurrently running transactions in system, where V is a set of vertices T_1, T_2, …, T_n, denoting the transactions in the system, and E is a set of directed edges between transactions, denoting the precedence among them. An arc is drawn from T_i to T_j, T_i -> T_j, $1 \le i, j \le n, i \ne j$, if T_i read an item written by T_j, which has not committed yet, or T_j wrote an item (in its workspace) that has been read earlier by T_i.

Transactions in the system can be classified into 3 classes. A transaction that has not executed any conflicting operations is called an independent transaction. Once a transaction has executed its first conflicting operation, it becomes a preceding or preceded transaction, depending upon whether it precedes or is preceded by another transaction. To prevent the precedence graph from growing rampantly, once a transaction has become a preceding (or preceded) transaction, it shall remain a preceding (or a preceded) transaction for its entire life time.

Let T_i and T_j be two transactions that involve in a conflict operation. Regardless the conflict being RAW or WAR, let T_i be the transaction that performs a read on the item, while T_j the transaction that performs a write on that item. The conflict operation is allowed to proceed only if the following rule, called the Prudent Precedence Rule, is satisfied.

Prudent Precedence Rule:

T_i is allowed to precede T_j or T_j is allowed to be preceded by T_i if

(i) T_i has not been preceded by any transaction and

(ii) T_j has not preceded any other transaction.

We shall use Figure 1 to explain the properties of the resulting precedence graph for transactions following the Prudent Precedence Rule. It can be observed that the first condition of the rule (denoted by (i) in the rule) states that a preceded transaction cannot precede any transaction, as illustrated by the red arcs, marked with x, T7 to T1 and T3 to T4, in the figure, while the second condition (denoted (ii)) states that a preceding transaction cannot be preceded, as

illustrated by the red arcs, marked with x, T1 to T2 and T7 to T1, in the figure. Since there cannot be any arcs between nodes in the same class and there is no arc from the preceded class to the preceding class, the graph cannot have a cycle.

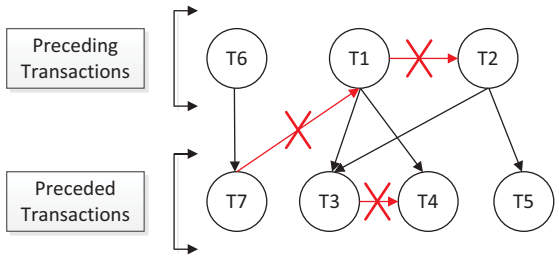


Figure 1. The Precedence Graph

2.3. Prudent Precedence Protocol

Each transaction is executed in three phases: read, wait-to-commit, and commit phases. In the read phase, transactions proceed following the precedence rule. Once a transaction finishes all its operations, it enters the wait-to-commit phase, waiting for its turn to commit following the precedence established in the read phase. Transactions release resources in the commit phase. In the following, we describe in details each phase.

2.3.1. Read Phase

A transaction executing a conflict operation with another transaction will be allowed to proceed if it satisfies the prudent precedence rules; otherwise, it will be either blocked or aborted. The transaction that violates the precedence rules is hereafter called a violating transaction.

In the following, we show a situation with a violating transaction.

Example 3. There are three transactions. Their operations and schedule are as follows.
T_1: $R_1(b)$ $W_1(a)$ ...
T_2: $R_2(a)$ $W_2(e)$ ...
T_3: $R_3(e)$ …
Schedule: R_1 (b) $W_1(a)$ $R_2(a)$ $W_2(e)$ ~~$R_3(e)$~~

T_2 -> T_1 is established when T_2 reads "a", and T_2 becomes a preceding transaction. Later when T_3 tries to read "e" ($R_3(e)$), the operation is suspended (denoted by ~~$R_3(e)$~~ in the schedule) because T_2, a preceding transaction, cannot be preceded. Thus, T_3 becomes a violating transaction and needs to be blocked or aborted.

The simplest strategy to handle a violating transaction, such as T_3, is to abort it. Unfortunately, aborts may waste the efforts already spent. Therefore, we prefer blocking with the hope that the violation may later resolve and the violating transaction T_3 can still complete later. For example, T_3 is blocked, i.e., $R_3(e)$ is postponed; if T_2 eventually commits, then T_3 can resume and read the new value of "e" produced by T_2. The read/write with the Prudent Precedence Rule is summarized in Figure 2.

```
if there is a RAW or WAR conflict
{
    if the prudent precedence rule is satisfied,
        proceed with the operation;
    else
        abort or block;
}
```

Figure 2. Read/Write with Prudent Precedence Rule

Let us elaborate on the blocking of a violating transaction a bit. By allowing a violating transaction to block, a transaction can now either be in an active (or running) state or a blocked state. Although blocking can increase the survival rate of a violating transaction, it can also hold data items accessed by the violating transaction unused for extended periods. Therefore, a time quantum must be set up to limit the amount of time a violating transaction can wait (block itself), just like the 2PL. Once the time quantum expires, the blocked (violating) transaction will be aborted to avoid building a long chain of blocked transactions.

Theorem 1. The precedence graph generated by transactions following the Prudent Precedence Rule is acyclic.

Proof. As explained in Section 2.2, there cannot be a cycle in the precedence graph following the Prudent Precedence Rule. As for violating transactions, they will either abort by timeouts or resume executions if the violation disappear due to the aborts or commits of the other transactions with which the transactions conflict. In either case, it does not generate any arcs that violate the Prudent Precedence Rule, and the graph remains acyclic.

2.3.2 Wait-to-Commit Phase

Once a transaction finishes its read phase, it enters the wait-to-commit phase, waiting for its turn to commit because transactions may finish the read phase out of the precedence order established.

First, each transaction acquires exclusive locks on those items it has written in the read phase to avoid building further dependencies. Any transaction in the read phase wishes to access a locked item shall be blocked. If such a blocked transaction already preceded a wait-to-commit transaction, it shall be aborted immediately in order not to produce a circular wait, that is, wait-to-commit transactions wait for their preceding blocked transactions to complete or vice versa. Otherwise, the blocked transaction remains blocked until the locked item is unlocked. Figure 3 shows the locking when a transaction accesses a locked item.

A transaction can proceed to the commit phase if no transactions, either in the read or the wait-to-commit phase, precede it. Otherwise, it has to wait until all its preceding transactions commit.

```
/* Ti is accessing an item x

  if x is locked
{       if x is locked by a transaction preceded by Ti
            abort Ti;
        else
            block Ti (until x is unlocked);
}
read/write with the Prudent Precedence Rule (Figure 2);
```

Figure 3. Accessing Locked Items

2.3.3 Commit Phase

As soon as a transaction enters the commit phase, it flushes updated items to the database, releases the exclusive locks on data items obtained in the wait-to-commit phase, and also releases transactions blocked by it due to violations of the precedence rule. Figure 3 summarizes the wait-to-commit and the commit phases.

```
/* when a trans. Ti reaches its wait-to-commit phase */
Wait-to-Commit Phase:
  Lock items written by Ti;
  Ti waits until all preceding transactions have committed;

Commit Phase:
  Flush updated items to database;
  Release locks;
  Release transactions blocked by Ti;
```

Figure 4. Wait-to-Commit and Commit Phases

Example 4. Suppose that we have the following transactions, T_1, T_2:

T_1: $R_1(a)$, $R_1(b)$

T_2: $R_2(b)$, $W_2(a)$, $W_2(b)$

Assume that the following is the schedule:

$R_1(a)$, $R_2(b)$, $W_2(a)$, $W_2(b)$, [wc_2], ~~$R_1(b)$~~ $abort_1$, wc_2, c_2

When T2 writes "a" (W2(a)), T1 ->T2 is established, due to an earlier R1(a). So, when T2 reaches its wait-to-commit phase, denoted by wc2, it locks both "a" and "b". However, T2 has to wait until T1 has committed, denoted by [wc2], due to the established precedence T1 ->T2. Later, when T1 tries to read "b", it is aborted, as indicated by R1(b) and abort1, because "b" is locked by T2, as stipulated in Figure 3. Now no transaction is ahead of T2, so it can finish its wait phase (wc2) and commits (c2).

2.4. Serializability

A history is a partial order of the operations that represents the execution of a set of transactions [5]. Let H denote a history. The serialization graph for H, denoted by SG_H, is a directed graph whose nodes are committed transactions in H and whose edges are $T_i \rightarrow T_j$ ($i \neq j$) if there exists a T_i's operation precedes and conflicts with a T_j's operation in H. To prove that a history H is serializable, we only have to prove that SG_H is acyclic.

THEOREM 2. Every history generated by the Prudent Precedence Protocol is serializable.

Proof. The precedence graph is acyclic as proved in Theorem 1. The wait-to-commit phase enforces the order established in the precedence graph to commit. So, the serialization graph has no cycle and is serializable.

3. SIMULATION RESULTS

This section reports the performance evaluation of 2PL, OCC, and the Prudent Precedence Concurrency Control (PPCC) by simulations.

3.1 Simulation Model

We have implemented 2PL, OCC and PPCC in a simulation model that is similar to [1]. Each transaction has a randomized sequence of read and write operations, with each of them separated by a random period of a CPU burst of 15±5 time units on average. All writes are performed on items that have already been read in the same transactions. All writes are stored in private work space and will only be written to the database after commits following the strict protocol.

3.2 Parameter Settings

Our goal is observe the performance of the algorithms under data contentions. The write operations cause conflicts and thus the data contentions. Therefore, we shall experiment with different write probabilities, 20% (moderate), and 50% (the highest), to observe how the two algorithms adapt to conflicts. Other factors that affect the data contentions are database sizes and transaction sizes. Therefore, two database sizes of 100 and 500 items, and two transaction sizes of averaged 8 and 16 operations will be used in the simulation.

Table 3.1: Parameter Settings

Database size	100, 500 items
Average transaction size	8± 4, 16±4 operations
Write probability	20%, 50%
Num. of CPUs	4, 16
CPU burst	15±5 time units

Transactions may be blocked in 2PL, OCC and PPCC to avoid generating cycles in the precedence graphs. Blocked transactions are aborted if they have been blocked longer than specified periods. We have experimented with several block periods and select the best ones to use in the simulations.

The primary performance metric is the system throughput, which is the number of transactions committed during the period of the simulation. This is an overall performance metric.

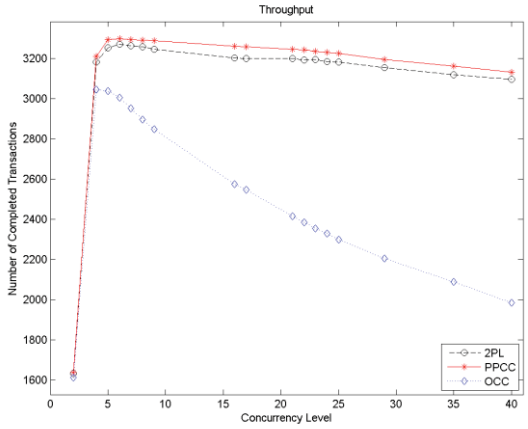


Figure 4(a). DB size 500

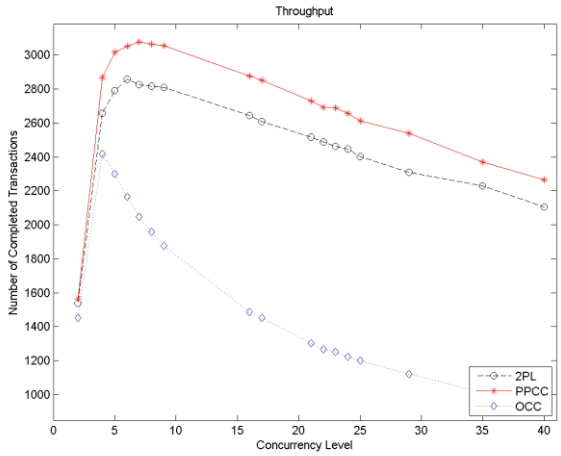


Figure 4(b). DB size 100

Figure 4. Write probability 0.2, Transaction Size 8

3.3 Experimental Results

In this section, we report the simulation results on the two protocols based on the above setups.

3.3.1 Data Contention

As mentioned earlier, the data contention is mainly caused by the write operations. If transactions have no writes, there will be no conflicts and all three protocols will have identical performance.

Given the same write probability, the greater the transaction sizes, the greater the numbers of write operations are in the system, and thus the higher the data contentions are. On the other hand, given the same number of write operations, the smaller the database size, the greater the chance of conflicts, and thus the higher the data contentions are. Here, we will see how these factors affect the performance of the three protocols.

We experimented with two database sizes, 100 items and 500 items, and two transaction sizes, averaged 8 and 16 operations in each transaction. The experimental results in this subsection were obtained with the setup of 4 CPUs. The simulation time for each experiment is 100,000 time units.

- **Write probability 0.2**

Given the write probability 0.2, each transaction has on average one write operation for every four reads.

Figures 4 shows the performance for transactions with averaged 8 (8 ± 4) operations for two databases of sizes 500 (Figure 4(a)) and 100 (Figure 4 (b)). As observed, as the level of concurrency increased initially, the throughput increased. At low concurrency levels, all protocols had similar throughputs because there were few conflicts. But as the concurrency level increased further, conflicts or data contention intensified and the increase in throughput slowed down a bit. After a particular point, each protocol reached its peak performance and started to drop, known as *thrashing*.

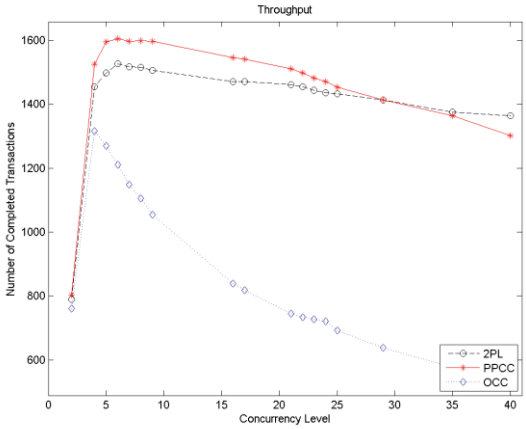


Figure 5(a). DB size 500

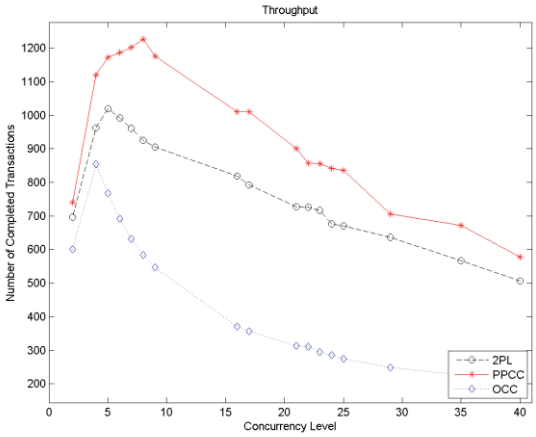


Figure 5(b). DB size 100

Figure 5. Write probability 0.2, Transaction Size 16

For database size 500 (Figure 4(a)), the highest numbers of transactions completed in the given 100,000 time unit period were 3,299 for PPCC, 3,271 for 2PL, and 3,046 for OCC, that is, a 0.86% and 8.31% improvements over 2PL and OCC, respectively.

In Figure 4(b), the database size was reduced to 100 items to observe the performance of these protocols in a high data contention environment. The highest numbers of completed transactions were 3,078, 2,857, and 2,417 for PPCC, 2PL, and OCC, respectively, i.e., an 7.74% and 27.35% higher throughputs than 2PL and OCC. This indicates that PPCC is more effective in high data contention environments than in low data contention environments, which is exactly the purpose that we design the PPCC for.

Now, we increase the average number of operations in each transaction to 16 while maintaining the same write probability 0.2. Figure 5 shows the results.

For database size 500 (Figure 5(a)), the highest throughput obtained by PPCC was 1,605, while 2PL peaked at 1,527 and OCC at 1,316. PPCC had a 5.11% and 21.96% higher throughputs than 2PL and OCC. As for database size 100 (Figure 5(b)), the highest throughputs obtained were 1,226, 1,019, and 854 for PPCC, 2PL and OCC, respectively. PPCC had a 20.31% and 43.56% higher throughputs than 2PL and OCC.

In general, as the data contention intensifies, PPCC has greater improvements over 2PL and OCC in performance.

- **Write probability 0.5**

With the write probability 0.5, every item read in a transaction is later written too in that transaction. Figure 6 shows the throughput of the two protocols with the average number of operations set to 8 per transaction.

The highest numbers of transactions completed during the simulation period (Figure 6(a)) were 3,258 for PPCC, 3,237 for 2PL, and 2,978 for OCC for database size 500, a slight improvement over 2PL(0.65%), but a much larger improvement over OCC (9.40%). As the database size decreased to 100 (Figure 6(b)), the highest numbers of completed transactions were 2,898, 2,803, and 2,365 for PPCC, 2PL, and OCC, respectively, that is, a 3.39% and 22.54% higher throughput than 2PL and OCC, due to the higher data contentions.

Figure 7 shows the throughput of the three protocols with the number of operations per transaction increased to 16.

The highest numbers of transactions completed during the simulation period (Figure 7(a)) were 1,490 for PPCC, 1,480 for 2PL, and 1,213 for OCC for database size 500, a 0.68% and 22.84% improvements over 2PL and OCC. As the database size decreased to 100 (Figure 7(b)), the highest

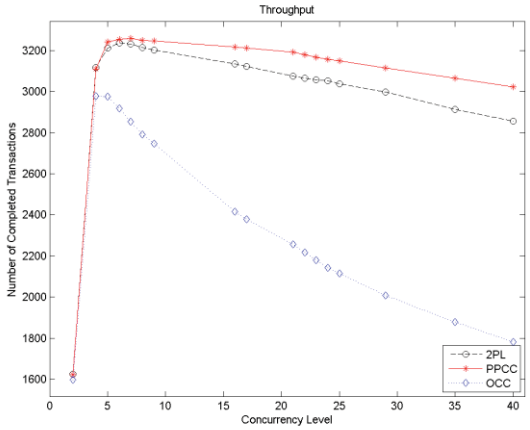


Figure 6(a). DB size 500

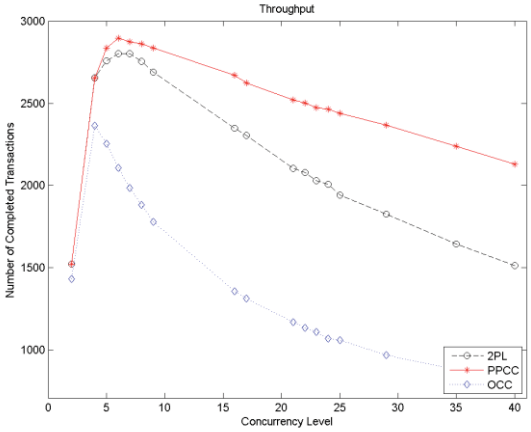


Figure 6(b). DB size 100

Figure 6. Write probability 0.5, Transaction Size 8

numbers of completed transactions were 1,011, 969, 747 for PPCC, 2PL, and OCC, respectively, that is, a 4.33% and 35.34% higher throughputs than 2PL and OCC.

In very high data contention environments, few transactions can succeed, as illustrated in Figure 7(b). This indicates that there is still room for improvement in designing a more aggressive protocol that allows more concurrent schedule to complete serializably.

4. CONCLUSIONS

The proposed protocol can resolve the conflicts successfully to a certain degree. It performed better than 2PL and OCC in all situations. It has the best performance when conflicts are not very severe, for example, in situations where transactions are not very long and write probabilities are not too high. Further research is still needed for resolving more complex conflicts while keeping the protocols simple.

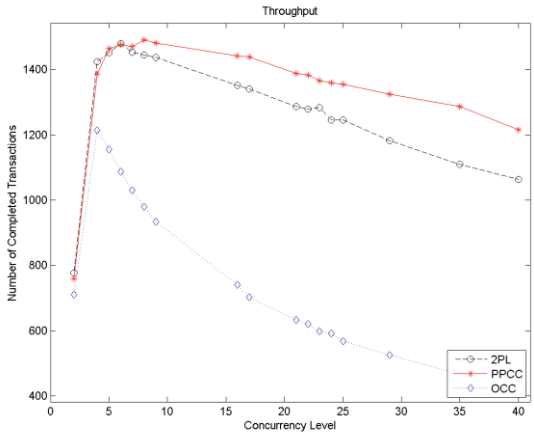


Figure 7(a). DB size 500

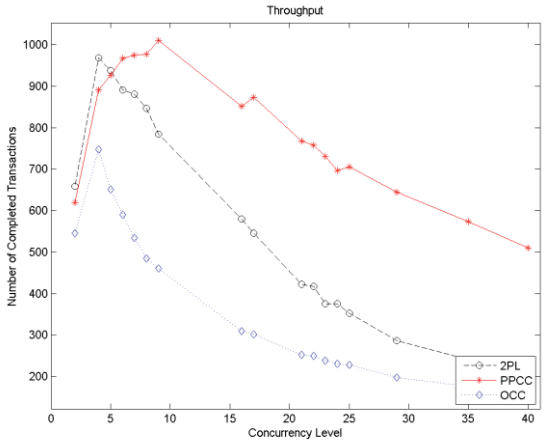


Figure 7(b). DB size 100

Figure 7. Write probability 0.5, Transaction Size 16

REFERENCES

[1] R. Agrawal, M. J. Carey, and M. Livny, "Concurrency Control Performance Modeling: Alternatives and Implications," ACM Transactions on Database Systems, 12(4), pp. 609-654, 1987.

[2] P. A. Bernstein, V. Hadzilacos, and N. Goodman. (1987) Concurrency Control and Recovery in Database Systems. Addison-Wesley, Reading, MA.

[3] P. Bernstein, N. Goodman, "Timestamp-based Algorithm for Concurrency Control in Distributed Database Systems", Proc. VLDB 1980, pp. 285 – 300.

[4] P. Bernstein, N. Goodman, J. Rothnie, Jr., C. Papadimitriou, "Analysis of Serializability in SDD-1: a System of Distributed Databases", IEEE Transaction on Software Engineering SE-4:3, 1978, pp. 154 -168.

[5] M. Carey, M. Livny, "Distributed Concurrency Control Performance: A Study of Algorithms, Distribution, and Replication", Proc. 14th VLDB Conference, pp. 13-25, 1988.

[6] S. Ceri, S. Owicki, "On The Use Of Optimistic Methods for Concurrency Control in Distributed Databases", Proc. 6th Berkeley Workshop, pp. 117-130, 1982.

[7] P. Eswaran , J. N. Gray , R. A. Lorie , I. L. Traiger, (1976) The Notions of Consistency and Predicate Locks in a Database System, Communications of the ACM, Vol. 19, No. 11, p.624-633.

[8] T. Haerder, (1984) Observations on Optimistic Concurrency Control Schemes. Information Syst., 9, 111-120.

[9] J. R. Haritsa, , M. J. Carey, and M. Livny, (1990) Dynamic Real-Time Optimistic Concurrency Control. In Proc. 11th Real-Time Systems Symp., Lake Buena Vista, FL,5-7 December, pp. 94-103.

[10] H. Kung, and J. Robinson, (1981) On Optimistic Methods for Concurrency Control. ACM Trans. Database Syst., 6, 213-226.

[11] K.W. Lam, K.Y. Lam and S.L. Hung. Distributed Real-time Optimistic Concurrency Control Protocol. In Proceedings of International Workshop on Parallel and Distributed Real-time Systems, Hawaii, pp. 122-125, IEEE Computer Society Press (1996).

[12] S. Mullender and A. S. Tanenbaum. "A Distributed File Service Based on Optimistic Concurrency Control," Proc. 10th ACM Symp. On Operating System Principles, 1985, pp. 51-62.

[13] D. Reed, "Implementing Atomic Actions on Decentralized Data", ACM Transactions on Computer Systems, 1,1, pp. 3-23, 1983.

[14] I. Ryu, A. Thomasian, "Performance Evaluation of Centralized Databases with Optimistic Concurrency Control", Performance Eval. U, 3, 195, 211, 1987.

[15] Özgür, U and Alejandro, B. "Exploiting main memory DBMS features to improve real-time concurrency control protocols." ACM SIGMOD Record 25.1 (1996): 23-25.

Mobile Application for Interpreting Telugu Language

Mark Smith
Department of Compute Science, University of Central Arkansas, Conway, AR, USA

Abstract – *The usage of mobile devices has increased dramatically in recent years. These devices serve us in many practical ways and provide us with many* applications*. One of the more* recent advances is the use of language/*text translation into speech. A complete mobile application (i.e., App) is described in this work for translating English text into Telugu and Telugu text into English by using a standard Google Language API. The Telugu text is subsequently converted into synthesized speech by utilizing the popular Festival Speech synthesis system that supports a number of languages including Telugu. Many papers have cited the problems of synthesizing Telugu speech accurately. A novel algorithm is introduced (runs offline of the Mobile App) that measures the accuracy and understandability of the Festival system. A sample of the Telugu language is extracted from the speech of native speakers (and stored in a standard database) and is compared to the speech generated by the Festival Speech synthesis system. The algorithm utilizes the MPEG-7 audio descriptors as the core level features extracted from both the generated and the synthesized speech. A distance measurement is performed and the word with the best measurement is selected using a quick-lookup mechanism. that the most appropriate video was selected. The completed App and measurement results are provided.*

Keywords: Speech Synthesis, Intelligent Systems

1 Introduction

Mobile computing has become one of the fastest evolving areas of computer science. Consumer demand and interest in mobile devices has exploded, as exemplified by the introduction of smart phones such as Apple's iPhone [4], T-Mobile's Android [2,12], and Microsoft Windows Phone[1]. Indeed, many programmers have found new opportunities developing applications (better known as *Apps*) for these smart phones as exhibited by almost 2 billion apps downloaded for the iPhone to date. One popular App involves the language translation from one language to another. Many of the more commonly spoke languages (Spanish, French, German, Chinese, etc.) have been successfully synthesized by a variety of systems [1,5,7], while other languages are still maturing in this area and are a prime area of research. This work explores the interpretation of the Telugu language text into synthesized speech. Telugu is the second most commonly spoke language in India and consists of an intricate alphabet with many complicated symbols (approximately 60) while presenting numerous challenges to speech synthesis. This paper describes a mobile App focused on the Telugu language consisting of the following features:

1. Translation of Telugu text to English
2. Translation of English text to Telugu
3. Translation of English text to English speech
4. Translation of Telugu text to Telugu speech

The first 3 steps are accommodated by the popular Google AJAX Language API [2,4] and are easily integrated into the App as standard features. The 4th feature is a much less defined feature and has proven to be a prime area of research. This paper describes the first 3 steps in detail within the initial sections. The last feature also has a measurement applied to it in an attempt in quantifying the quality of the speech synthesis. The popular Festival speech synthesizer is utilized in this system and is used for generating a variety of commonly used words in the Telugu language. The synthesized words are compared with sound files created by native speakers speaking the same words as those that were synthesized. The spoken works are stored in a MySQL database and are compared to the synthesized words by using the MPEG-7 Audio Descriptors. The algorithm is summarized by the following steps:

1. MPEG-7 Audio Signature, Spectrum Centroid, Spectrum Spread, and Spectrum Envelope are extracted from a series of spoken words performed by native speakers. The sound file and the corresponding MPEG-7 features are stored in a MySQL database.
2. The same MPEG-7 descriptor features is extracted from synthesized Telugu speech (based on text) .
3. The MPEG-7 features are matched based on the a distance measurement and a voting algorithm. If both the Audio Signature Spectrum and the Audio Envelope descriptor have a distance measurement is within range and at least one is the minimal possible, the audio file is selected as the best matching.

Step 3 is repeated for approximately 100 words and the results are tabulated in the last section

2 Festival Speech Synthesis

The primary objective of a text to speech processing system is to convert typed text representing individual words (in this case the text is entered into spoken words. This is significantly different from converting English text to Telugu text by which this conversion is quite easily performed by the

Google language APIs. For example English text to Telugu text can be described by the following diagram:

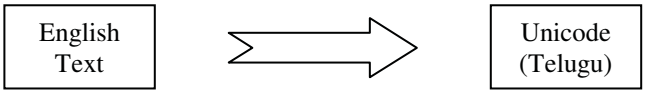


Fig. 1 English to Telugu (Unicode)

An example of converting the English greeting "Hello" to Unicode and subsequently into Telugu is shown below:

English	Telugu
Hello	మహాభారతం

Fig. 2 English to Telugu Example

Where each Telugu character in the right column is created as a series of Unicode characters as shown below:

0c2e -------మ

0c39 -------హ

0c3e ------ -ా

Fig. 3 Telugu Unicode Example

A more challenging aspect for this system is to however convert the actual text to the spoken word. Much work has been performed over the last several years in the way of speech synthesis [2,3,6]. A very popular speech synthesis system that has been utilized in many iOS Apps as well as desktop applications has been the much heralded Festival speech synthesis and Festox speech voice box. The Festival system was developed by the University of Edinburgh with major contributions made by the well known research institute Carnegie Mellon. It offers a full text to speech system with various APIs, as well as an environment for development and research of speech synthesis techniques. It is written in C++ with a Scheme-like command interpreter for general customization and extension[9]. Festival is designed to support multiple languages, and comes with support for English (British and American pronunciation), Welsh, and Spanish. Voice packages exist for several other languages, such as Castilian Spanish, Czech, Finnish, Hindi, Italian, Marathi, Polish, Russian and Telugu. Festival provided the speech synthesis portion of our system necessary for our application to convert text to spoken Telugu speech [16]. The compete App using the Festival speech synthesis system is explained in detail in Section 5.

Our system provided the best speech synthesis as possible as provided by the Festival speech conversion server. But we were concerned about the accuracy of the Telugu speech provided by Festival. The next sections describes our approach in testing Festival by using the MPEG-7 audio descriptors as outlined in [15]. The descriptors are used in comparing the generated speech created by Festival with spoken words generated by actual Telugu speakers. The spoken words from the actual Telugu speakers have the same attributes extracted from them as the synthesized words from the Festival speech synthesized system. The attributes then undergo a distance measurement used for comparing the closeness of the synthesized with the spoken words. These steps are performed for each of the MPEG-7 Audio descriptor. A voting algorithm is then used in classifying the Festival synthesized word as sufficiently matching the actual spoken word from a real Telugu speaker.

3 MPEG-7 Audio Descriptors

MPEG-7 Audio provides structures—in conjunction with the Multimedia Description Schemes part of the standard—for describing audio content. Utilizing those structures are a set of low-level Descriptors, for audio features that cut across many applications (e.g., spectral, parametric, and temporal features of a signal), and high-level Description Tools that are more specific to a set of applications. Those high-level tools include general sound recognition and indexing Description Tools, instrumental timbre Description Tools, spoken content Description Tools, an audio signature Description Scheme, and melodic Description Tools to facilitate query-by-humming. The MPEG-7 audio descriptors provide the primary set of attributes for comparing streaming or recorded speech. The most powerful audio features as described by [14] are:

1. Audio Signature
2. Envelope Signature
3. Spectrum Centroid

A detailed description for each of these descriptors is provided in the subsequent sections.

4 Audio Signature

While low-level audio Descriptors in general can serve many conceivable applications, the spectral flatness Descriptor specifically supports the functionality of robust matching of audio signals. The Descriptor is statistically summarized in the Audio-Signature-Description-Scheme as a condensed representation of an audio signal designed to provide a unique content identifier for the purpose of robust automatic identification of audio signals. Applications include audio fingerprinting, identification of audio based on a database of known works and, thus, locating metadata for legacy audio content without metadata annotation. The Audio Signature is by far the simplest of all the audio descriptors to interpret, understand, and utilize in equations that utilize some type of distance measurement. The audio signature utilizes 43 specialized attributes that are very useful for retrieving and comparing audio data[13].

5 Envelope Signature

The basic spectral audio Descriptors all share a common basis, all deriving from a single time-frequency analysis of an audio signal. They are all informed by the first Descriptor, the Audio-Spectrum-Envelope Descriptor, which is a logarithmic-frequency spectrum, spaced by a power-of-two divisor or multiple of an octave. This Audio Spectrum Envelope is a vector that describes the short-term power spectrum of an audio signal. It may be used to display a spectrogram, to synthesize a crude auralization of the data, or as a general-purpose descriptor for search and comparison. The envelope signature is a much lower level descriptor and requires more analysis in utilizing this descriptor in a comparison operator. An example of data generated by the 57 features created by the envelope signature is shown below in Fig. 4:

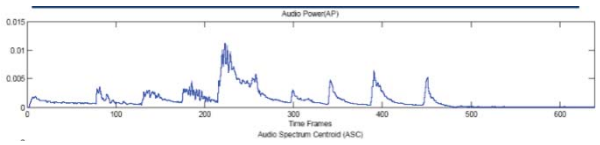


Fig. 4. Envelope Signature

6 Spectrum Centroid

The Spectral-Centroid-Descriptor is the power-weighted average of the frequency of the bins in the linear power spectrum. As such, it is very similar to the Audio-Spectrum-Centroid Descriptor, but specialized for use in distinguishing musical instrument timbres. It is has a high correlation with the perceptual feature of the "sharpness" of a sound. The four remaining timbral spectral Descriptors operate on the harmonic regularly-spaced components of signals. For this reason, the descriptors are computed in linear-frequency space. The Harmonic-Spectral-Centroid is the amplitude-weighted mean of the harmonic peaks of the spectrum. It has a similar semantic to the other centroid Descriptors, but applies only to the harmonic (non-noise) parts of the musical tone[11]. An example of a plot of the Spectrum Centroid applied to a commonly spoken Telugu greeting is shown below in Fig. 5:

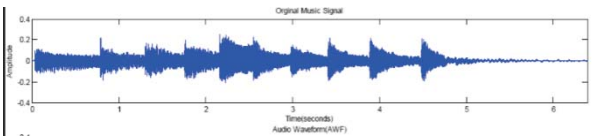


Fig. 5. Spectrum Centroid

7 Feature Extraction

A set of spoken words from actual Telugu speakers is stored in the form of MP3 sound files in a standard MySQL database management system. Approximately 100 such spoken words are stored as Binary Large Objects (i.e., BLOBs) in a column of the MySQL database. The three MPEG-7 audio features described in sections 4 through 7 are extracted as shown below using equations (1) through (3):

$$\overline{x}_{as} = [a_0, a_i ... a_{N-1}] \quad (1)$$

$$\overline{x}_{es} = [a_0, a_i ... a_{M-1}] \quad (2)$$

$$\overline{x}_{ec} = [a_0, a_i ... a_{P-1}] \quad (3)$$

Where $\overline{x}_{as}$ represents the feature vector corresponding to the MPEG-7 audio signature descriptor, $\overline{x}_{es}$ represents the feature vector corresponding to the MPEG-7 envelope signature descriptor, x_{ec} represents the feature vector corresponding to the MPEG-7 spectrum centroid descriptor, a_i represents the *ith* attribute extracted from the specified descriptor, *N* is 32, *M* is 43 and *N* is 57. Each spoken word from the actual Telugu speaker has these 3 sets of feature vectors extracted and pre-stored with the MP3 file implemented as a Binary Large Object. At real-time, the same three feature vectors are extracted from the Festival speech synthesis and compared with the corresponding stored vectors[10]. The subsequent sections describes the process for matching and classifying the Festival synthesized word with the best matching spoken word.

8 Word Classification Process

Next, the process to select a set of candidate Telugu speaker audio files based on the Festival synthesized word is presented. The main concept is to first find a set of candidates which are closest to the synthesized word. This filtering step will next be followed a classification step that will find the best matching word best on this described filtering process. First, the distance between the Festival spoken word and each stored word is performed for each feature vector. The average of all such distances is given below in equation (4):

$$\overline{\mu} = \frac{\sum_{j}^{S} \sum_{i}^{N} \sqrt{(a_{if} - a_{ijs})^2}}{S-1} \quad (4)$$

Where a_{if} is the *ith* attribute of a given feature vector for the synthesized word, a_{ijs} is the *ith* attribute for the *jth* sample of a given feature vector for a given spoken word, and $\overline{\mu}$ is the mean vector of all distance differences computed between all samples the spoken word and the synthesized words. The standard deviation for these differences are then computed as:

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N-1}(D_i - \mu)^2}{N-1}} \quad (5)$$

Where *Di* is a difference as computed in (4), μ is the average difference as computed in (4), and σ is the standard deviation of the difference as computed between the Festival synthesized word and a given spoken work for a given feature vector. The standard deviation is used for computing an adaptive threshold for a given feature vector as shown in (6):

$$T_{thresh} = \mu + 2\sigma \quad (6)$$

Where T_{thresh} is the adaptive threshold as computed for a given feature vector and σ and μ are the same as before. The adaptive threshold is computed for all three feature vectors and are all combined linearly and weighted the same as shown in Equation (7):

$$T_{tot} = wT_{as} + wT_{es} + wT_{ec} \quad (7)$$

Where *w* is assigned 1/3. This guarantees that all thresholds have the same weight. All differences Di that is less than Ttot as given in (8)

$$D_i < T_{tot} \quad (8)$$

are selected as the initial candidates for the best matching audio file for the actual Telugu speaker and the synthesized Festival generated word. Additional sets of candidates are selected on those words that are less than each individual threshold:

$$D_i < T_{es|as|ec} \quad (9)$$

The candidates are then filtered based on those words which are found in each set of candidates based on (8) and (9) – i.e., the intersection of each the candidate sets as given by (10):

$$C_{tot} \cap C_{as} \cap C_{es} \cap C_{ec} \quad (10)$$

The candidate word of (10) that minimizes the *Di* distance is then chosen as the best matching word between the synthesized and spoken word. The results will show that (10) provides the best set of candidates and will therefore be the criteria most used when selecting the spoken word that best matches the synthesized word from Festival The next section discusses the iOS App implementation.

9 iOS App Implementation

A complete iOS application was implemented on the Mac OS platform using XCode 7.0, Swift 2.0, and iOS 9.1. The initial screen that is displayed when the app starts is shown below in Fig. 6:

Fig. 6 Initial App Screen

The use has the option for entering either English or Telugu in the From/To and the translation will occur. Telugu language has a special alphabet that consists of approximately 60 characters that can be entered from a special keyboard customized for the language characters. Examples of the characters that can be entered from the keyboard is shown below if Fig. 7:

Fig. 7 Telugu Keyboard Example

Note the sound indicator that allows the user to hear the translated word along with the text translation or to simply view the text translation by itself. A complete English-Telugu dictionary is built into the App that allows the user to view meanings in English or Telugu as shown below in Fig. 8:

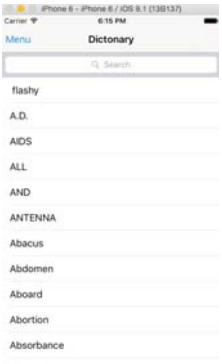


Fig. 8 Dictionary Example

In addition, a history of all Telugu/English words is maintained so the user can re-select and reuse as needed. This

is a convenience feature that users can find helpful when repeated translations. An example of the history page of the App is shown below in Fig. 9:

Fig. 9 History of Telugu Words

The App was tested thoroughly on the 100 sample words used in testing the synthesized words generated by Festival. The results for each candidate group and the overall results of the best matching word is shown in the next section containing all results for the iOS App built in this system.

10 Results

A fully function iPhone App implemented in section 9 was fully tested for all 100 sample words stored from actual Telugu speakers. The testing involves examining if the synthesized word correctly chooses the actual spoken word from each of the different MPEG-7 descriptors. of the testing involves determining the number of incidents correctly identified in contrast to those incidents falsely detected. Table 1 shown below illustrates the results of the system

Table 1 Results

Descriptor	T1	T2	T3	T4
Audio Signature	77	82	91	98
Envelope Spectrum	53	77	72	89
Envelope Centroid	73	72	77	91
All Descriptors	70	92	97	98

Where T1, T2, T3, and T4 are the combination of the various thresholds computed in equations (6) – (9). T1 is the threshold computed for Audio Signature, T2 is the threshold computed for Envelope Spectrum, T3 is the threshold for Envelope Centroid, and T4 is the combination of all 3 thresholds as given in (7). The best combination of the descriptors appears to be the Audio Signature while Envelope Spectrum does not seem to be quite as accurate. The numbers represent the percentage correct for each synthesized word.

The results appear very promising illustrating the accuracy of this system. The error rate is well within bounds and provides users with a very accurate speech translation app.

11 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.

Data Cleaning in Out-of-Core Column-Store Databases: An Index-Based Approach

Feng Yu[1], **Tyler Matacic**[1], **Weidong Xiong**[2], **Mohammed Ali A Hamdi**[2], **Wen-Chi Hou**[2]
[1]Computer Science and Information Systems, Youngstown State University, Youngstown, OH, USA
[2]Department of Computer Science, Southern Illinois University, Carbondale, IL, USA

Abstract

Write optimization in out-of-core (or external memory) column-store databases is a well-known challenge. Timestamped Binary Association Table (or TBAT) and Asynchronous Out-of-Core Update (or AOC Update) have shown significant improvements for this problem. However, after a time period of AOC updates, the selection query performance on TBAT gradually decreases. Even though data cleaning methods can merge update records in TBAT to increase ad-hoc searching speed, it could be a time-consuming process. In this work, we introduce multiple data cleaning methods utilizing the index structure called *offset B^+-tree* (or *OB-tree*). When the OB-tree and updating records can be fit into the system memory, an *eager data cleaning* approach is introduced for fast cleaning speed. In a data intensive environment, the OB-tree index or the updating records might be too large to fit into memory; therefore, a *progressive data cleaning* approach is introduced which can divide the update records into small slips and clean the data a memory-economic manner.

keywords: column-store database, data cleaning, index, B^+-tree

1. Introduction

Column-store databases (also known as columnar databases or column-oriented databases) have drawn much attention recently. They refer to the databases that vertically partition data and separately store each column. The history of column-store databases can be traced back to 1970s when transposed files were implemented in the early development of DBMS, followed by applying vertical partitioning as a technique of table attribute clustering. By the mid-1980s, the advantage of a fully *decomposed storage model* (*DSM*) over the traditional row-based storage model (*NSM* or *Normalized Storage Model*) was studied [1]–[3].

TAXIR (TAXonomic Information Retrieval) is the first automatic application of column-store database focusing on biological information retrieval and management [4], [5]. KDB and Sybase IQ were the first two commercially available column-store databases developed in 93 and 95, respectively. It's not until about 2005 when many open-source and commercial implementations of column-store databases took off [6]. The well-known column-store databases include: Apache Cassandra [7], Apache HBase [8], MonetDB [9], KDB, SAP HANA [10], [11], and Vertica [12].

The data storage in a column-store database is vertically partitioned and sharded by projecting each column into a separate fragment. A vertical fragment is referred as a BAT (Binary Association Table) [9], which is stored contiguously on a large enough disk page in order to mitigate seeking overheads across multiple ranges of data. The data in each BAT is densely patched in order to improve I/O performance, and also rapidly compressed utilizing light-weight compression schema to improve storage efficiency.

One of the benefits of a column-store database is its information retrieval speed, which is much faster than a row based database. Thanks to the DSM feature, the column-store database fits well into the write-once-and-read-many environment. The column-store database works especially well for OLAP and data mining queries that retrieve a large number of tuples but only considers a small collection of attributes. Put simply, it can retrieve only the attributes included in the query prediction without the need to read the entire tuple. Another featured benefit of the column-store database is data compression, which can reach a higher compression rate and higher speed than traditional row-based databases. One of the major reasons for this higher compression is that the information entropy in the data of a single column is lower than that of traditional row-based data.

Optimizing write operations in a column-store database has always been a challenge. Existing works focus on write optimizations in a main-memory column-store database. Krueger et al. [13], [14] introduced the differential update to improve the write performance in MonetDB. A special columnar data structure called the *delta buffer* was introduced to temporarily store decomposed row-based input data. However, to the best of our knowledge, very few works focused on optimizing the write performance on the *out-of-core* (*OOC* or *external memory*) column-store databases.

Vertica [12], a column-store database for large vol-

ISBN: 1-60132-441-3, CSREA Press ©

id	name	balance
1	Alissa	100.00
2	Bob	200.00
3	Charles	300.00

(a) Row-Based Table `customer`

oid	int
101	1
102	2
103	3

(b) BAT `customer_id`

oid	varchar
101	Alissa
102	Bob
103	Charles

(c) BAT `customer_name`

oid	float
101	100.00
102	200.00
103	300.00

(d) BAT `customer_balance`

Fig. 1: `customer` Data in Row-Based and Column-Store (BAT) Format

ume OOC storage, introduces a specially designed data storage procedure called *k-safety* to ensure ACID of update transactions on large volumes of data and improve the data importation efficiency. Nevertheless, k-safety focuses more on the transaction control rather than the write performance improvement for high velocity update query streams.

In [15], an efficient solution was proposed to optimize the write operations (update and deletion) on an OOC column-store database. An operation called *Asynchronous Out-of-Core Update* (or *AOC Update*) was originally designed based on a new data structure called *Timestamped Binary Association Table* (or *TBAT*). There is a potential problem that, after a period of time of AOC updates, the selection query performance on TBAT gradually deteriorates [16].

In order to address the problem of performance deterioration after multiple AOC updates, data cleaning methods [15] can be used to clean up update records from the body of the TBAT into the appendix of the TBAT. In [16], online data cleaning methods are introduced in order to clean up TBAT without the need of locking the file. However, the data cleaning procedure can be time-consuming on large data sets [17].

In [18], a new index structure called *Offset B^+-tree* (or *OB-tree*) is introduced for fast data retrieval in the TBAT file. OB-tree is a succinct sparse index specially designed for a TBAT where AOC updates are performed. It replaces the global pointers with relative pointers, called *offset*, to save storage space. In addition, OB-tree supports fast searching queries including ad-hoc and range queries on TBAT.

In this research, we aim to introduce data cleaning methods utilizing the OB-tree index to achieve a higher speed performance. Based on the concurrent usage of a TBAT file while data cleaning is performed, we divide the TBAT conditions into *cold data* and *hot data* conditions. The cold data condition is when no other users could possibly change the TBAT while data cleaning is performed. However, updates could happen when data cleaning is performed simultaneously. In the interest of simplicity, this work focuses on the data cleaning algorithms on cold data. Furthermore, to adapt our research for a data intensive environment, we develop OB-tree based cleaning methods based on the data size compared with the memory size. When the OB-tree and the updating records can be fit into the system memory, we introduce an *eager data cleaning* approach for fast data cleaning. On the other hand, for memory bottleneck scenarios, we introduce the *progressive data cleaning* approach.

The rest of the paper is structured as follows. Section 2 is the background introduction of column-store databases. Section 3 shows the performance degeneration after AOC updates. Section 4 introduces data cleaning methods without using index. The OB-tree data structure is revisited in section 5. In section 6, we introduce multiple data cleaning methods utilizing the OB-tree index. The preliminary experiment results are shown in section 7. Section 8 is the conclusion and future works.

2. Inside Column-Store Databases

The data structure of a column-store database exclusively uses *BAT*s (*Binary Association Tables*). A BAT is a fragment of an attribute in the original row-based storage. It usually consists of an `oid` (Object Identifier) or `ROWID`, along with a column of attribute values, which in a pair is called a *BUN* (*Binary UNit*s). It is a physical model in a column-store database and the sole bulk data structure it implements. The BAT is categorized in a special group of storage models called *Decomposed Storage Model* (or *DSM*) [1], [2].

The row-based storage data is the original user input data, called the *front-end* data or *logical* data. To input the data into a column-store database, a mapping rule should be defined from the logical data structure to the physical data structure, namely BAT.

Example 1. (From Row-Based Table to BAT) Suppose a row-based table is `customer`. It consists of three attributes `id`, `name`, `balance`, and `id` the primary key. The row-based data is shown in Fig. 1(a). In a columnar database, this logical table will be decomposed into 3 BATs namely `customer_id`, `customer_name`, `customer_balance`. Each BAT contains two columns: an oid and an attribute value column with the column name as the corresponding column data type.

In Example 1, the logical table is fully decomposed into 3 BATs, Fig 1(b)-1(d), with each BAT containing one of the attributes. This is also referred to as *full*

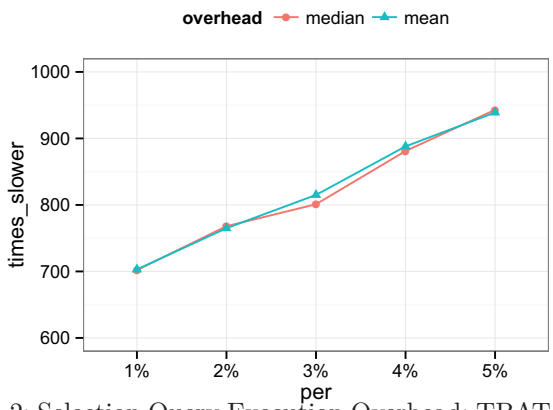


Fig. 2: Selection Query Execution Overhead: TBAT over BAT ($\frac{\text{time(TBAT)}}{\text{time(BAT)}} \times 100\%$)

vertical fragmentation [6]. Full vertical fragmentation has many advantages. First of all, data accessing is efficient for queries accessing many rows but with fewer columns involved in the query. Another advantage is the reduction of the workload on the CPU and memory generated by OLAP and data mining queries, which typically consider only a few columns in a logical table.

Compared to fully vertical fragmentation, the other pattern is *partial vertical fragmentation* [19]. It assumes the prior knowledge of which columns are frequently accessed together. Also, it employs the attribute usage matrix to determine optimal clustering of columns into vertical fragments. However, OLAP and data mining are application areas that indicate ad-hoc queries, as a good OLAP or data mining system must be able to quickly answer queries involving attributes of arbitrary combinations. Nevertheless, the partial vertical fragmentation is useful to detect the data block location in a distributed database system.

3. Selection Speed Degeneration after AOC Updates

The AOC update [15] is a fast update method on column-store database. It will increasingly create pending data in the appendix of the TBAT. For small TBAT files and a small amount of AOC updates, the effect on selection queries can hardly be noticeable. However, for larger files and a large portion of AOC updates, the decreased selection speed cannot be ignored. The overhead comparing the selection execution time on TBAT vs BAT is illustrated in Figure 2. Using a randomly generated 1MB TBAT file, we perform AOC updates by 1% to 5% of the original file. We perform randomly generated ad hoc selection queries, with a selection ratio of 10%, on both normally updated BAT files and AOC updated TBAT files. The mean overhead and median overhead are 819% and 822%, respectively.

4. Data Cleaning without an Index

To improve the selection performance on the TBAT after AOC updates, multiple data cleaning methods which don't use indexes are developed in [15] and [16]. The purpose of data cleaning in an OOC column-store database is to detect the latest version of updated data and merge them into the body of the TBAT. By the requirement of locking the database, those methods are divided into two groups, namely *offline data cleaning* [15] and *online data cleaning* [16].

4.1 Offline Data Cleaning

Offline data cleaning can only be performed after the database has been locked to avoid inconsistant data during the cleaning process. An offline data cleaning method is introduced in [15], `merge_update` which can remove the duplicated TBUNs in the TBAT with same `oid` but different `timestamp`'s.

Offline data cleaning first employs a merge sort on the entire TBAT file including the body and appendix, and then deletes the duplicated TBUN's in a sequential manner. This requires profound time in execution when a large amount of AOC updates are accumulated. Even though the time issue is not the first concern for an offline data cleaning approach, a better time efficient manner is obviously preferred.

4.2 Online Data Cleaning

The online data cleaning approaches are developed in [16], which consist of an eager approach and a progressive approach for speed-priority and memory-priority, respectively.

The central idea of online data cleaning, compared with the offline approach, is to enable the users to continue querying the TBAT during the data cleaning procedure time. This is a major focus of the online approach especially when the database is a streaming environment, where the input is non-stop. The main difference for online data cleaning is the employment of a sophisticated data structure called a *snapshot*.

The online approach will first make a snapshot of the body and create a new appendix file linked to the TBAT. The older version of the appendix will be merged into the snapshot of the body utilizing merge sorting and binary searching. During this time, the TBUN's in the appendix will be written to the body as the traditional update on the BAT. After the merging is complete, the snapshot of the body will replace the original body in the TBAT, and the older version of appendix will be purged.

In a data intensive environment, the updated data might be too large to be fit into the main memory. Thus we separate the online data cleaning process into two different approaches, namely an eager approach and

a progressive approach, for speed priority and memory priority, respectively.

4.2.1 Online Eager Data Cleaning

The central idea of online eager data cleaning is to increase the merging speed. The entire appendix of the TBAT will be read into the memory and perform online merging into a snapshot of the body. Put simply, this approach merges the entire appendix at once. After that, the merged snapshot will replace the original body in the TBAT file.

4.2.2 Online Progressive Data Cleaning

Online progressive data cleaning fits for the data intensive scenarios where the entire appendix may not fit into memory. In these cases, the eager approach cannot be applied. The key concept in online progressive data cleaning is the appendix queue, where each TBAT can contain more than one appendix. The size of each appendix, or *block size*, needs to be manually defined by the database administrator, which cannot exceed the size of the available memory on the system. The original appendix of a TBAT file will then be split into separate appendixes according to the block size. The appendix queue will be attached to the TBAT instead of a single appendix.

During the progressive data cleaning procedure, each time an appendix is retrieved from the appendix queue, an eager data cleaning approach is then performed to merge the split appendix with the snapshot of the body. Simultaneously, the appendix queue can create a new split appendix file to accept streaming updates and enqueue the appendix once its size researches the block size.

5. Offset B^+-Tree Index

5.1 Data Structure

Offset B^+-tree, or simply *OB-tree* is introduced in [18]. Namely, the OB-tree is a variant of B^+-tree. It is developed based on the B^+-tree and has several important properties that the TBAT requires.

1. OB-tree has a succinct data structure that can be easily adopted by existing column-store databases.
2. An OB-tree is a sparse index for only the updated records in a TBAT. An OB-tree can be either stored in main memory or serialized on hard disk. When the OB-tree is stored in main memory, the data retrieving speed can be orders of magnitude faster.
3. OB-tree allows the insertion of duplicated keys. In fact, the key in an OB-tree is the `oid` in the TBAT data file. It is possible that a record associated with one `oid` is updated multiple times.

Inside an OB-tree, there are two categories of nodes, namely internal nodes and leaf nodes. The top node is the root node. It is a leaf node when the OB-tree has only one layer, and an internal node when the OB-tree has multiple layers. Associated with each OB-tree, the parameter n determines the layout of every node inside the OB-tree.

- Each internal node will have space for n search keys, i.e. `oid`s, and $n+1$ pointers that can be used to point to other nodes. At least, $\lceil (n+1)/2 \rceil$ of the pointers are used. If the internal node is a root, we only require that at least 2 pointers are used regardless of the value of n.
- Each leaf node will also have space for n search keys. But among the $n+1$ space units, only 1 allocation is used to point to the next leaf node in the sequence. The left n units are reserved to save a special value, called *offset*, used to point to the location of the updated record in the appendix of a TBAT.
- How to assign each `oid` to each node is the same as in a B^+-tree.

An *offset* in the OB-tree is a scalar recording of the relative location of an updated record inside an appendix that is appended at the end of the body of TBAT. We assume the number of lines of the body is l_b and the number of lines of the appendix is l_a. For any given record located at the kth line, $1 \leq k \leq l_a$, of the appendix, the *offset associate with this updated record* is k. Intuitively, the updated record with offset k is at the $(l_a + k)$th line in the entire TBAT.

Remark 1. The offset is a scalar pointer to the target record in the TBAT. The space cost of an offset can be relatively smaller than a pointer in any operating system. The employment of the offset can be considered as a simplified method of pointer elimination.

Since pointers can occupy additional spaces inside any B/B^+-tree, we use this scalar offset to replace most of the pointers at the leaf nodes, and the actual location of a record can be promptly calculated by the definition of offset. The data type of an offset can be flexible and decided by the user. Typically, it can be the same data type of the `oid`, which can be a 4 byte or 8 byte integer. To save more space, smaller bytes of an integer can also be considered with the assumption that at most a smaller portion of the TBAT records are updated.

Example 2 (A Basic Example of OB-Tree)**.** We demonstrate a simple OB-tree. We set the parameter n to be 3, i.e. each node in the OB-tree can store up to 3 `oid`s and 4 pointers. From the appendix, we can generate `oid`-offset pairs of (`oid`, offset), indicating the `oid`s and the associated offsets. Suppose the given `oid`-offset pairs are $\{(1,1),(2,2),...,(10,10)\}$. Figure 3 demonstrates the OB-tree after inserting each `oid` and offset.

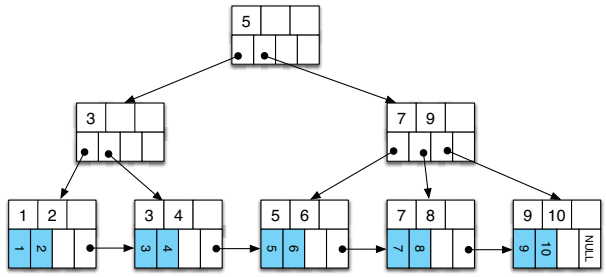


Fig. 3: An Example of OB-Tree

Algorithm 1 OB-Tree Eager Cleaning on Cold Data

Input: tbat: TBAT file; obtree: OB-tree
Output: tbat: TBAT file after data cleaning

```
1: procedure OB-Tree-Clean-Eager(tbat, obtree)
2:     n_update ← update_list.getLength( )
3:     update_list ← obtree.outputUpdateList( )
4:     n_body ← tbat.body_length
5:     row_num_1 ← 1
6:     while update_list.hasNext( ) do
7:         ⟨oid, offset⟩ ← update_list.getNext( )
8:         updated_record ← tbat.readLine(n_body + offset)
9:         row_num_2 ← tbat.body.binarySearch(oid, row_num_1, n_body)   ▷ binary search starting from row_num_1 and return found location
10:        tbat.body.seekRelative(row_num_2-row_num_1)
11:        tbat.body.writeAtCurrentLine(updated_record)
12:        row_num_1 ← row_num_2   ▷ use last found locatoin as the next start
13:    end while
14:    tbat.appendix.destroy( )   ▷ destroy appendix after cleaning
15:    tbat.close( )   ▷ close TBAT file
16:    return SUCCESS
17: end procedure
```

The numbers in the upper part of each node denote the `oid`s inserted. The internal nodes resemble the internal nodes inside a B$^+$-tree. However, the leaf nodes are different in the pointer locations. The blue blocks in 3 denote the parts where offsets are stored in replacement of pointers. Only the last pointer in each leaf nodes is retained pointing to the next leaf node in the sequence. In addition, the end pointer in the last leaf node is a `NULL` pointer.

6. Data Cleaning using OB-Tree

In this section, we discuss how to use the OB-tree for data cleaning on TBAT files with updated records. The main purpose of data cleaning on TBAT files is to merge the updated records from the appendix part of the TBAT into the body. Because of the unsorted nature of appendix records, where the updated records may not necessarily be sorted according to `oid`'s, searching on the appendix is a time-consuming process.

Algorithm 2 OB-Tree Progressive Cleaning on Cold Data

Input: tbat: TBAT file; update_slip_queue: the queue of updating slips from OB-tree
Output: tbat: TBAT file after data cleaning

```
1: procedure OB-Tree-Clean-Progressive(tbat, update_slip_queue)
2:     while update_slip_queue.hasNext( ) do
3:         update_slip←update_slip_queue.dequeue( )
4:         OB-Tree-Clean-Eager(tbat, update_slip)
5:     end while
6: end procedure
```

The major advantage of using an OB-tree for data cleaning is to hasten the searching speed when looking for those updated records and their locations. The `oid`'s are well organized in the OB-tree, and their locations can be easily calculated by retrieving their offsets associated with `oid`'s in the OB-tree. Compared with data cleaning without any index [15], [16], OB-tree-based data cleaning features a fast retrieval speed on updated records because of its B$^+$-tree nature. In addition, there is no need to pre-sort the appendix of the TBAT which also saves significant system time.

Because the OB-tree is a variant of a B$^+$-tree, range searchings for `oid`'s are rather fast, we can output from the OB-tree the updated `oid`'s with their offsets into an update list and perform merging while reading this update list. Based on the size of the updated records in the appendix, one can choose to use either an *eager cleaning* approach when the update list can be fit into the main-memory, or a *progressive cleaning* approach when the update list is too large to be fit into the main-memory. Please note that in the latter situation, OB-tree itself could also be too big to be fit into the main-memory; thus, we need to serialize the OB-tree and progressively retrieve updated `oid`'s so that the entire procedure can be processed.

Data cleaning on the TBAT can happen on both *cold data* and *hot data*. Cold data refers to the senario when the TBAT is in a locked condition and no writing from other users is allowed. On the other hand, hot data means there could be other users writing to the TBAT file while data cleaning is in execution. In this work, we focus on the senario of cold data cleaning since data cleaning on hot data is discussed in [16].

6.1 Eager Data Cleaning

If the OB-tree and the update list of $\langle \text{oid}, \text{offset} \rangle$ can be fit into memory, an eager data cleaning approach is preferred. In addition, we assume the `oid`'s in the TBAT body may not necessarily be consecutive. The algorithm of eager cleaning is illustrated in Algorithm 1.

In the first phase, we generate the update list of $\langle \text{oid}, \text{offset} \rangle$ that need to be cleaned from the OB-tree. Since the OB-tree is a B^+-tree variant, one can first search for the left-most leaf node, and then start to find its sibling nodes until all leaf nodes are exported. The reason that we export all updating `oid`'s and their offsets at once, instead of one-by-one, is to avoid multiple searches in the OB-tree which could be time-consuming. The update list is currently sorted by `oid` in ascending order.

In the second phase, we retrieve from the update list the records which need to be merged into the body. Each time, we first retrieve a pair of $\langle \text{oid}, \text{offset} \rangle$ from the update list. The appended record with the latest value is located at the line (tbat.body_length + offset). The target record, which needs to be updated in the body of the TBAT, can be searched using a special deductive binary searching method. It can be located by a binary search with the row number range from the previous target row number, or row_num_1, to the maximum row number in the body, or n_{body}. Especially, when searching for the first `oid` in the update list, we let row_num_1 equal to 1. In such a manner, the next binary search can use the previous binary search result to lower the time cost. This deductive searching approach ends when the pairs in the update list are all merged into the TBAT body. After all of this is completed, the appendix of the TBAT file can be destroyed.

6.2 Progressive Data Cleaning

If the OB-tree or the update list of $\langle \text{oid}, \text{offset} \rangle$ is too large to be fit into the memory, a progressive data cleaning approach is preferred. In this scenario, we assume the OB-tree is serialized into secondary storage and read into the memory by segmentations. The algorithm of progressive data cleaning is illustrated in Algorithm 2.

The key difference of the progressive cleaning approach is to organize all updating pairs of $\langle \text{oid}, \text{offset} \rangle$ into slips, called update slips. The size of the update slips can be manually determined by the administrator according to the hardware and operating system configurations. For each update slip read into memory, we can use the eager approach to merge the pairs of $\langle \text{oid}, \text{offset} \rangle$ into the TBAT body. Please note that, when merging each update slip, there is no need to read the OB-tree because the offset information is included into the update slip. Therefore, the data cleaning speed is guaranteed.

7. Data Cleaning Experiments

In these experiments, we perform comprehensive comparison of data cleaning with OB-tree (the index-based approach) and without any help of index (the traditional approach). Since our goal is to test the performance improvement using the OB-tree index, we assume all the updated data is loaded into the memory at once. In the traditional approach, the updated records are first sorted in-memory, and then merged into the TBAT body by using binary searching. The only memory usages for all methods include the storage for OB-tree and the sorting for updated records.

7.1 Experiment Design

We focus on synthetic TBAT datasets of 64MB to mimic a data block in HDFS (Hadoop Distributed File Systems) [20]. We generate random queries to update the datasets with 5 updating ratios from 1% to 3% with the increment of 5%. After that, both data cleaning approaches are performed on same updated datasets to compare their performances in speed and space costs.

7.2 Result Analysis

Three key measurements are recorded during the tests including execution time, disk access count, and memory cost of OB-tree.

First of all, Figure 4 depicts the results of data cleaning execution times on a 64MB dataset using OB-tree and traditional methods. It is obvious that on each stage of update ratio the index-based approach is faster than the traditional approach. The speed difference between the two methods become more obvious with the increment of the updating ratio and the size of data. This can also be observed in Figure 5, where the improvement is measured by relative overhead between two methods defined as

$$\left(\frac{\text{time(traditional)} - \text{time(OB-Tree)}}{\text{time(OB-Tree)}} \right) \times 100\%$$

The borderline lies on the updating ratio of 20% where their differences exceed 10%. In addition, the performance difference is increasing steadily with the updating ratio.

In general, the OB-tree index prominently improves the data cleaning process in all tests compared with the traditional method without using any index. The memory cost associated with OB-tree is minor and increases slowly with the updating ratio. This memory cost can be further mitigated by properly adjusting the parameters of the OB-tree.

8. Conclusion and Future Works

In this research, we introduce data cleaning methods on the TBAT using the OB-tree index. We introduce

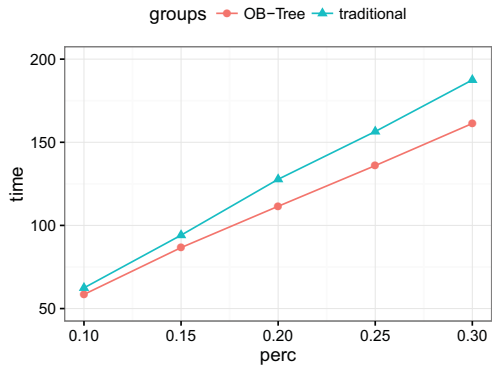


Fig. 4: Data Cleaning Execution Time (sec) on 64MB Dataset

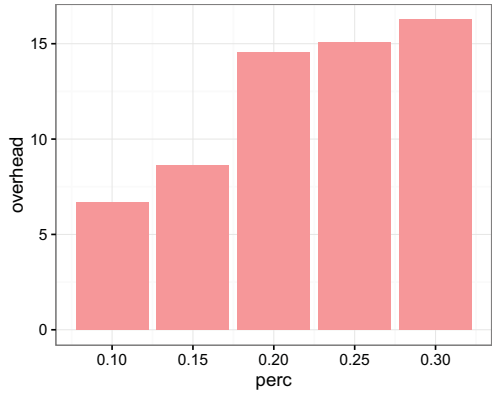


Fig. 5: Data Cleaning Execution Overhead (%) on 64MB Dataset

two data cleaning approaches, namely the eager cleaning approach and the progressive cleaning approach. The eager cleaning approach assumes the data size can be fit into memory. On the other hand, the progressive cleaning approach is designed for a large dataset scenario. Preliminary experiments have shown that the proposed data cleaning methods utilizing OB-tree is much faster than traditional methods.

For future work, we will continue to perform comparison experiments on big data sets. More extensive experiments will be designed to compare their speed performance and space costs.

References

[1] G. P. Copeland and S. N. Khoshafian, "A decomposition storage model," in *ACM SIGMOD Record*, vol. 14, no. 4. ACM, 1985, pp. 268–279.

[2] S. K. G. C. T. Jagodits and H. B. P. Valduriez, "A query processing strategy for the decomposed storage model," in *Proceedings*. Order from IEEE Computer Society, 1987, p. 636.

[3] M. Zukowski, N. Nes, and P. Boncz, "DSM vs. NSM: CPU performance tradeoffs in block-oriented query processing," ser. DaMoN '08. New York, NY, USA: ACM, 2008, pp. 47–54.

[4] R. Brill, *The Taxir Primer*. ERIC, 1971.

[5] G. F. Estabrook and R. C. Brill, "The theory of the TAXIR accessioner," *Mathematical Biosciences*, vol. 5, no. 3, pp. 327–340, 1969.

[6] D. J. Abadi, P. A. Boncz, and S. Harizopoulos, "Column-oriented database systems," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1664–1665, aug 2009.

[7] G. Ladwig and A. Harth, "CumulusRDF: Linked data management on nested key-value stores," in *The 7th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2011)*, 2011, p. 30.

[8] L. George, *HBase: The Definitive Guide*. O'Reilly Media, Inc., 2011.

[9] P. Boncz, T. Grust, M. Van Keulen, S. Manegold, J. Rittinger, and J. Teubner, "MonetDB/XQuery: a fast XQuery processor powered by a relational engine," ser. ACM SIGMOD 2006, 2006, pp. 479–490.

[10] F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner, "SAP HANA Database: Data Management for Modern Business Applications," *SIGMOD Rec.*, vol. 40, no. 4, pp. 45–51, Jan. 2012.

[11] F. Färber, N. May, W. Lehner, P. Große, I. Müller, H. Rauhe, and J. Dees, "The SAP HANA database – an architecture overview." *IEEE Data Eng. Bull.*, vol. 35, no. 1, pp. 28–33, 2012.

[12] A. Lamb, M. Fuller, R. Varadarajan, N. Tran, B. Vandiver, L. Doshi, and C. Bear, "The vertica analytic database: C-store 7 years later," *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 1790–1801, Aug. 2012.

[13] J. Krueger, M. Grund, C. Tinnefeld, H. Plattner, A. Zeier, and F. Faerber, "Optimizing write performance for read optimized databases," ser. DASFAA'10. Tsukuba, Japan: Springer-Verlag, 2010, pp. 291–305.

[14] J. Krueger, C. Kim, M. Grund, N. Satish, D. Schwalb, J. Chhugani, H. Plattner, P. Dubey, and A. Zeier, "Fast updates on read-optimized databases using multi-core CPUs," *Proc. VLDB Endow.*, vol. 5, no. 1, pp. 61–72, Sep. 2011.

[15] F. Yu, C. Luo, W.-C. Hou, and E. S. Jones, "Asynchronous update on out-of-core column-store databases utilizing the timestamped binary association table," in *Proc. CAINE'14*, New Orleans, Louisiana, USA, 2014, pp. 215–220.

[16] ——, "Online data cleaning for out-of-core column-store databases with timestamped binary association tables," in *Proc. CATA'15*, Honolulu, Hawaii, USA, 2015, pp. 401–406.

[17] F. Yu and W.-C. Hou, "A framework of write optimization on read-optimized out-of-core column-store databases," in *Proc. DEXA 2015*, Valencia, Spain, 2015, pp. 155–169.

[18] F. Yu and E. S. Jones, "Hastening data retrieval on out-of-core column-store databases using offset B^+-tree," in *Proc. CAINE'15*, Diego/Harbor Island, San Diego, California, USA, 2015, pp. 313–318.

[19] D. Gluche, T. Grust, C. Mainberger, and M. Scholl, "Incremental updates for materialized OQL views," in *Deductive and Object-Oriented Databases*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1997, vol. 1341, pp. 52–66.

[20] T. White, *Hadoop: The Definitive Guide*. O'Reilly, March 2015.

Libwebex: A Non-discriminating Web Content Extraction Library

Michael Gruesen, En Cheng
Department of Computer Science
College of Arts and Sciences
University of Akron
Akron, OH 44325-4003

Abstract – Libwebex is designed to be a powerful and flexible tool for the purpose of web page content extraction in the domain of Data Integration. Utilizing the semi-structured nature of web pages, this library provides users with a skeletal view of the web page content. The data elements that need to be extracted are arranged into a hierarchical tree of related topics and content. In this representation, users are able to build custom applications on top of this framework using the object-oriented design concept, the Visitor design pattern. All a programmer needs to provide is the customized behavior to be executed at each node in the tree. The library provides multiple basic building blocks for additional applications, such as archiving, searching, and printing mechanisms.

Keywords: Data Integration; Web Content Extraction; Web Page Interpreter

1 Introduction

In the domain of Data Integration, extracting data from semi-structured data sources, particularly web pages, presents numerous challenges. There exist many approaches to dealing with these types of data sources; however their usage is generally limited to the scope of the application. Generally this functionality is achieved by utilizing wrappers that extract certain information from explicit sources. Manual wrappers are the most powerful as their extraction rules are tailored directly to data sources. Learning based wrappers, such as HLRT [1] and Stalker [2], attempt to lighten the burden of the programmer by using preconfigured extraction rules. However, the common theme is that these wrappers require knowledge of how source lays out its information and what fields are desirable.

We propose a general purpose Web Extraction Library, *Libwebex*, as a tool to extract all possibly relevant information from web page sources for the purpose of Data Integration. To demonstrate the behavior and usage of this library we will use an example web page, Figure 1, from the Software Defined Networking (SDN) research group Flowgrammable [3], specifically a page that focuses on the OpenFlow [4] networking protocol. This web source contains a fairly large knowledge base that explains concepts related to this up-and-coming networking architecture paradigm. Building on the ideas implemented in HTML web page interpreters, *libwebex* gives the user a scaled down view of the web page, highlighted in Figure 1.

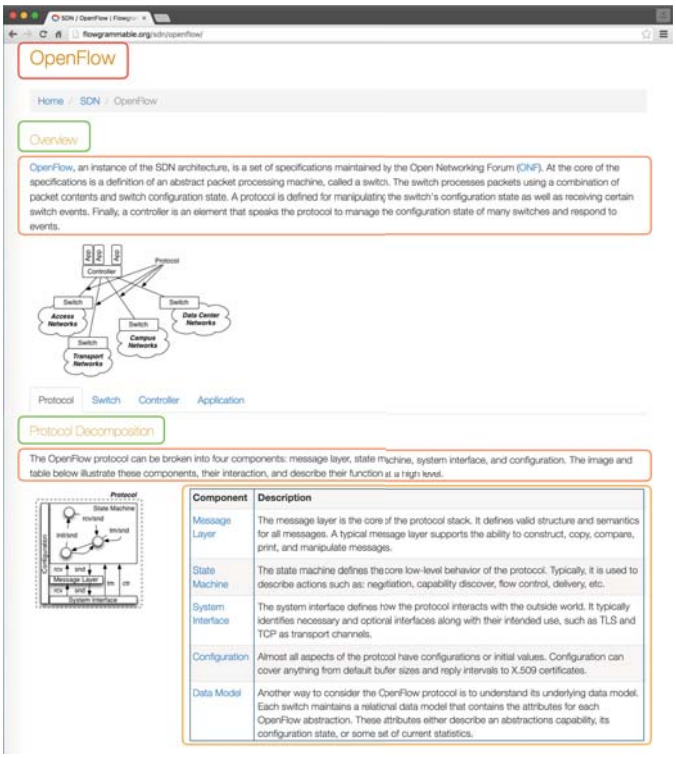


Figure 1. An example of web page source where highlighted areas are the extraction targets.

By utilizing the semi-structured nature of web pages based on the more relevant tags (e.g. title, headings, paragraphs, tables, and lists) expressed as a tree, additional applications can be easily built on top of this framework. Using some of the built-in behavior provided by the library, such as searching, users can create a more robust querying mechanism that utilizes a 3rd party string matching library to identify internal nodes, or topics, and return the directly related content associated with them. A more user-friendly, and aesthetically appealing, visualization component could be extending from the printing function provided. Lastly, integration with a 3rd party analytical engine that operates on tree-based data representations could also be achieved.

2 Libwebex

The web extraction library offers the ability to transform HTML web pages, laid out in a semi-structured fashion, into a tree of related topics and content. Each node in the tree represents either a section or content within the source, and the edges between them illustrate their relations. The distance between them helps show how closely related the two items are. Initially the lexer scans and forms target tag items from the source and adds them to a token list. Once this list has been formed, the different sections, subsections, and content are stitched into a hierarchical relation by the parser. Figure 2 shows how the input source is transformed through multiple stages and the output is a tree illustrating the relations between sections and content with the input source.

Libwebex is implemented using C++11 and currently targets Linux platforms. Thus the only dependencies for this library are a C++ 11 compliant compiler (e.g. GCC-4.8+). The following sections elaborate on the different parts of this library. Currently the user interface to the core *Libwebex* applications are command-line driven.

2.1 Targets

In order to identify the desired information from a web page, *libwebex* looks for predefined target tags that are considered to be relevant, eliminating non-relevant tags such as styling, scripting, and metadata. By utilizing the semi-structured tagged nature of web pages, the extraction rules are much more relaxed, allowing the user to focus on the actual content found rather than constructing extractors for particular data fields.

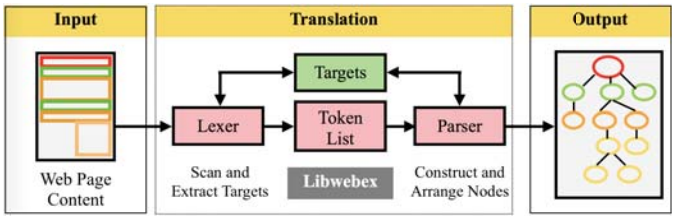


Figure 2. The *Libwebex* processing pipeline

To simplify the concepts of sections and content, we take an object-oriented approach w.r.t. their internal representations. Each tag found is considered either a section, establishing the scope of the current topic, or content. Section nodes can be the title and body tags or heading tags, while content nodes can be any of the supported content types, tables, lists, and text. There is a sound inheritance lineage between the content types, where tables are considered a container of multiple lists, and lists are containers of multiple text items. As such, a content type can be seen as either a container type, table or list, or a value type, text.

2.2 Lexing

As the first stage of the interpreter, the lexer identifies section and content tags supported by the library. The input source is scanned, stripping un-needed characters (e.g. '<', '!', '/'.) to form possible tokens. When a supported opening tag is found, it is added to the token list and the subsequent text is considered the value of that tag. A closing tag of the same type is used to delimit the end of a lexical token.

During the lexing phase, the first item that the lexer looks for is the 'title' tag, which serves as the main topic for the data source. Once found, the lexer will skip ahead until the 'body' tag is encountered, denoting the beginning of the sources usable data. As different section tags are discovered, their level is assigned based on their type. Title and body tags are fused into one source tag, whose level is always 0, representing the head of the tree. Subsequent heading tags are assigned a value based on their heading level value, i.e. $h_1 = 1$, $h_2 = 2$, … , $h_N = N$. While processing the body of the source, data related to tag definitions (e.g. names, classes, styling) is ignored.

In the case of nested tags, tokens will continue to be generated as long as the subsequent nested tag is currently supported. This is useful in the case of table content types. Tables present an interesting obstacle, as their orientation can be horizontal or vertical. In order to discern their layout, we look for emphasis tags (e.g. 'strong' or 'b'). This helps us establish which direction the data is laid out, and makes forming the table sub-types, lists, much easier.

2.3 Parsing

The second stage of the interpreter is the parser, which is responsible for creating nodes from the token list and arranging them into a hierarchical tree that illustrates the relations between sections and content within a web page. A simple page with a title and plain text, as well as more complex (e.g. nested and semi-structured) pages are both accounted for by the lexer.

At the top level of a parse tree lies the title, or the main subject of the source. After establishing the head, subsequent heading tokens are added to the tree using their precedence to determine their hierarchical arrangement. In the *Libwebex* interpreter, precedence is evaluated from least to greatest; a node with a lesser value should be placed above a node with greater value. This ensures that content types will always be the child of a section. During this process, the source and body nodes will be fused together, as a body token is an artifact from the lexing phase that has no meaning during the parsing phase.

Two headings of the same level are considered co-related sub-topics to their parent topic, while a heading of greater value is seen as a sub-topic of the current heading in scope. Sub-topics and related content are stored as a list of child nodes for each internal node. Figure 3 (A) shows a simplified example of what this translation process would produce, while Figures 3 (B) and (C) illustrate the extraction of container types, i.e. tables and lists.

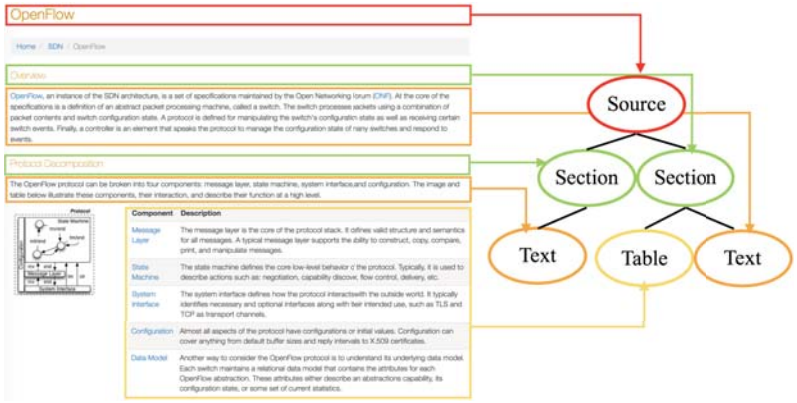


Figure 3. (a) An example for web data extraction

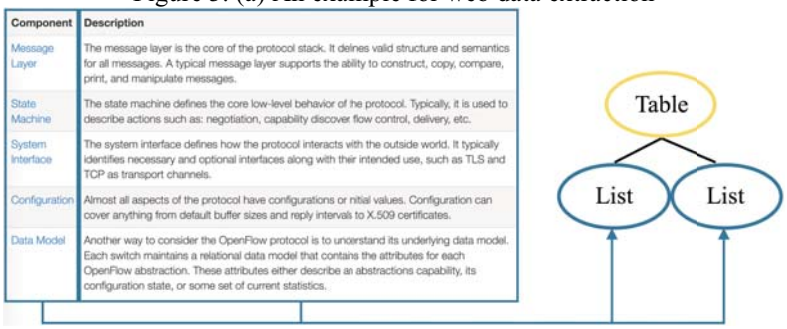


Figure 3. (b) Elaborated table extraction

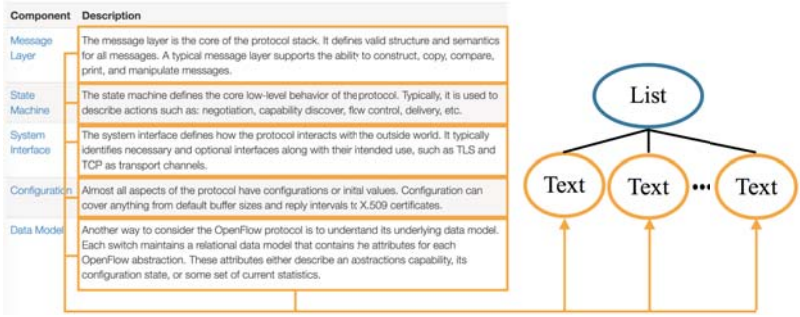


Figure 3. (c) Elaborated list extraction

2.4 Applications

After constructing a tree from the web page, the resulting structure allows for many interesting applications to be built on top of it utilizing the visitor pattern [5]. The visitor pattern is an object oriented design concept that dispatches "visitors" to the appropriately derived object when using inheritance, and is an alternative to using multiple pure virtual functions. This makes extending the library to support additional targets much simpler, as the user need not define multiple virtual functions for one type.

Each node in the tree 'accepts' a visitor object, who executes the desired behavior, defined by that visitor type, on the current node before (or after) moving on to the next. The visitor pattern is well suited for this structure as it allows for customized control over the tree traversal method, and eliminates the need to check if you are at a leaf node, rather the user defines the behavior at a leaf node.

The core applications provided in *libwebex* include archiving, searching, and printing, and are denoted by the prefix 'wbx_'. Custom applications can easily be built on top of the *libwebex* framework by implementing a visitor type over the supported section and content types, defining the desired behavior at each node, and linking against the library. These built-in applications also provide definitions for the corresponding visitor type, which users can utilize in their own applications. Below is an explanation of how these applications operate and their desired usage.

1) *Archive* – The archive application transforms a tree to Java Script Object Notation (JSON) [6] and writes the serialized string to file. This can be used as input to other *libwebex* applications. Archive files for *libwebex* are denoted by the '.wbxa' file extension.

2) *Search* – In the search application, the tree structure is searched for sections or content related to the given input. If found, the most directly related content (i.e. nodes directly descending from the selected one) will be printed to the output stream given.

3) *Print* – The print application directs the content from a given tree to the given output stream. The input can be a freshly parsed tree or an archived one.

3 Conclusions

The web extraction library, *libwebex*, is designed to be a flexible tool that is easily extended to fit a variety of users' needs. By utilizing a tree-based representation, it allows for the use of well-known tree algorithms to produce powerful and efficient applications. Its greatest strength is that it does not impose the burden of defining extraction rules to gather data from particulate web page sources, rather it pulls anything that could be relevant from the source.

4 Future Work

To make *libwebex* more robust, optimizations could be made to the lexing and parsing stages of the processing pipeline w.r.t. memory allocation. In addition to optimizations, the scope of targets could also be extended to support more content types (e.g. images and hyperlinks). Additional filtering of the content found by the lexer can be implemented by adding an elaboration phase to the parser. Multi-platform support would also broaden the appeal and usability of this library. Additionally, more core applications could be built into the *libwebex* framework to allow integration with 3rd party applications that utilize tree-based representations of data, as well as improved visualization output. Lastly, a graphical user interface would make using the library's core applications much more user friendly.

5 References

[1] Nicholas Kushmerick *et al*., "Wrapper Induction for Information Extraction," in *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*, Nagoya, Japan, 1997, pp. 729-736.

[2] Ion Muslea *et al*., "STALKER: Learning Extraction Rules for Semistructured Web-based Information Sources," Univ. of So. Cal., Los Angeles, CA, Tech. Rep. WS-98-14, 1998.

[3] Flowgrammable. SDN / OpenFlow [Online]. Available: http://www.flowgrammable.org/sdn/openflow.

[4] Open Networking Foundation. OpenFlow – Open Networking Foundation [Online]. Available: https://www.opennetworking.org/sdn-resources/openflow.

[5] Markus Schordan, "The Language of the Visitor Design Pattern," *J. Universal Comput. Sci.*, vol. 12, no. 7, pp. 846-867, Jul. 2006.

[6] *The JSON Data Interchange Format*. http://www.json.org/.

Using String Vector based KNN for Keyword Extraction

Taeho Jo
Department of Computer and Information Communication Engineering, Hongik University, Sejong, South Korea

Abstract— *In this research, we propose the string vector based KNN as the approach to the keyword extraction. The keyword extraction may be viewed as an instance of word classification, encoding words into numerical vectors may cause the main problems, such as the huge dimensionality, the sparse distribution and the poor transparency, and the problems were solved by encoding texts into string vectors in previous works on text mining tasks. In this research by these motivations, we encode words into string vectors, define the semantic operation on string vectors, and modify the K Nearest neighbor into its string vector based version which is used for the keyword extraction. As the benefits from this research, we expect the better performance and more compact representations than encoding words or texts into numerical vectors. Hence, the goal of this research is to implement the keyword extraction system with the benefits.*

Keywords: Keyword Extraction, String Vector, K Nearest Neighbor

1. Introduction

Keyword extraction refers to the process of extracting important words which are called keywords, from an article. The keywords are important indications for performing the information retrieval tasks, so we are interested very much in developing the schemes of extracting them. In this research, the keyword extraction is viewed into a binary word classification where each word is classified into a keyword or a non-keyword. We prepare the sample words which are labeled with 'keyword' or 'non-keyword', and construct the classification capacity by learning them. In this research, we assume that the supervised learning algorithms are used as the approach to the task, even if other types of approaches are available.

We mention some challenges with which this research attempts to tackle. In encodings texts or words into numerical vectors for using the traditional classification algorithms, many features are required, since each feature has very weak coverage[1]. Each numerical vector which represents a text or a word tends to be very sparse; it includes zero values dominantly[5][9]. Even if we proposed that texts or words should be encoded into tables in previous works, it was very expensive to compute the similarity between tables[5][9]. Therefore, in this research, we challenge against the above problems by encoding words into string vectors.

Let us consider some ideas which are proposed in this research. In this research, words are encoded into string vectors which consist of a finite ordered set of text identifiers as alternative representations to numerical vectors. We define the similarity measure between string vectors which is always given as a normalized value between zero and one; it corresponds to the cosine similarity between numerical vectors. The KNN (K Nearest Neighbor) is modified into the string vector based version, and applied to the special instance of word classification which is mapped from the keyword extraction. Note that in this research, the keyword extraction task is interpreted into the classification task.

Let us consider some benefits which are expected from this research. It is expected that string vectors are more compact representations of words which have much less features than numerical vectors. We expect the much better discriminations among string vectors than those among numerical vectors, since the sparse distributions can be avoided almost completely in each string vector. In this research, we expect also the improved performance by solving the above problems in encoding words into numerical vectors. Therefore, the goal of this research is to implement the keyword extraction systems which have the above benefits.

This article is organized into the four sections. In Section 2, we survey the relevant previous works. In Section 3, we describe in detail what we propose in this research. In Section 4, we mention the remaining tasks for doing the further research.

2. Previous Works

Let us survey the previous cases of encoding texts into structured forms for using the machine learning algorithms to text mining tasks. The three main problems, huge dimensionality, sparse distribution, and poor transparency, have existed inherently in encoding them into numerical vectors. In previous works, various schemes of preprocessing texts have been proposed, in order to solve the problems. In this survey, we focus on the process of encoding texts into alternative structured forms to numerical vectors. In other words, this section is intended to explore previous works on solutions to the problems.

Let us mention the popularity of encoding texts into numerical vectors, and the proposal and the application of string kernels as the solution to the above problems. In 2002, Sebastiani presented the numerical vectors are the standard

representations of texts in applying the machine learning algorithms to the text classifications [1]. In 2002, Lodhi et al. proposed the string kernel as a kernel function of raw texts in using the SVM (Support Vector Machine) to the text classification [2]. In 2004, Lesile et al. used the version of SVM which proposed by Lodhi et al. to the protein classification [3]. In 2004, Kate and Mooney used also the SVM version for classifying sentences by their meanings [4].

It was proposed that texts are encoded into tables instead of numerical vectors, as the solutions to the above problems. In 2008, Jo and Cho proposed the table matching algorithm as the approach to text classification [5]. In 2008, Jo applied also his proposed approach to the text clustering, as well as the text categorization [9]. In 2011, Jo described as the technique of automatic text classification in his patent document [7]. In 2015, Jo improved the table matching algorithm into its more stable version [8].

Previously, it was proposed that texts should be encoded into string vectors as other structured forms. In 2008, Jo modified the k means algorithm into the version which processes string vectors as the approach to the text clustering[9]. In 2010, Jo modified the two supervised learning algorithms, the KNN and the SVM, into the version as the improved approaches to the text classification [10]. In 2010, Jo proposed the unsupervised neural networks, called Neural Text Self Organizer, which receives the string vector as its input data [11]. In 2010, Jo applied the supervised neural networks, called Neural Text Categorizer, which gets a string vector as its input, as the approach to the text classification [12].

The above previous works proposed the string kernel as the kernel function of raw texts in the SVM, and tables and string vectors as representations of texts, in order to solve the problems. Because the string kernel takes very much computation time for computing their values, it was used for processing short strings or sentences rather than texts. In the previous works on encoding texts into tables, only table matching algorithm was proposed; there is no attempt to modify the machine algorithms into their table based version. In the previous works on encoding texts into string vectors, only frequency was considered for defining features of string vectors. In this research, based on [10], we consider the grammatical and posting relations between words and texts as well as the frequencies for defining the features of string vectors, and encode words into string vectors in this research.

3. Proposed Approach

This section is concerned with encoding words into string vectors, modifying the KNN (K Nearest Neighbor) into the string vector based version and applying it to the keyword extraction, and consists of the four sections. In Section 3.1, we deal with the process of encoding words into string vectors. In Section 3.2, we describe formally the similarity matrix and the semantic operation on string vectors. In Section 3.3, we do the string vector based KNN version as the approach to the keyword extraction. In Section 3.4, we focus on the process of applying the KNN to the given task with viewing it into a classification task.

3.1 Word Encoding

This section is concerned with the process of encoding words into string vectors. The three steps are involved in doing so, as illustrated in Figure 1. A single word is given as the input, and a string vector which consists of text identifiers is generated as the output. We need to prepare a corpus which is a collection of texts for encoding words. Therefore, in this section, we will describe each step of encoding the words.

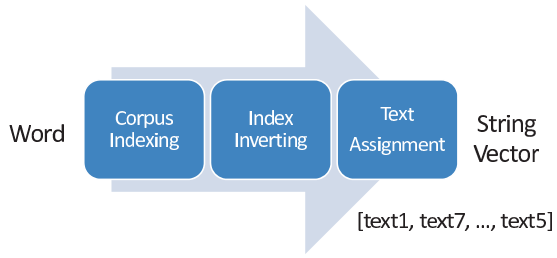


Fig. 1: Overall Process of Word Encoding

The first step of encoding words into string vectors is to index the corpus into a list of words. The texts in the corpus are concatenated into a single long string and it is tokenized into a list of tokens. Each token is transformed into its root form, using stemming rules. Among them, the stop words which are grammatical words such as propositions, conjunctions, and pronouns, irrelevant to text contents are removed for more efficiency. From the step, verbs, nouns, and adjectives are usually generated as the output.

The inverted list where each word is linked to the list of texts which include it is illustrated in Figure 2. A list of words is generated from a text collection by indexing each text. For each word, by retrieving texts which include it, the inverted list is constructed. A text and a word are associated with each other by a weight value as the relationship between them. The links of each word with a list of texts is opposite to those of each text with a list of words becomes the reason of call the list which is presented in Figure 2, inverted list.

Each word is represented into a string vector based on the inverted index which is shown in Figure 3. In this research, we define the features which are relations between texts and words as follows:

- Text identifier which has its highest frequency among the text collection
- Text identifier which has its highest TF-IDF weight among the text collection
- Text identifier which has its second highest frequency among the text collection

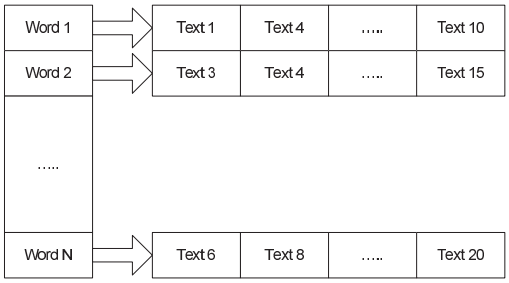


Fig. 2: The Inverted Index

- Text identifier which has its second highest TF-IDF weight among the text collection
- Text identifier which has its highest frequency in its first paragraph among text collection
- Text identifier which has its highest frequency in its last paragraph among text collection
- Text identifier which has its highest TF-IDF weight in its first paragraph among text collection
- Text identifier which has its highest TF-IDF weight in its last paragraph among text collection

We assume that each word is linked with texts including their own information: its frequencies and its weights in the linked texts and their first and last paragraphs. From the inverted index, we assign the corresponding values which are given as text identifiers to each feature. Therefore, the word is encoded into an eight dimensional string vector which consists of eight strings which indicate text identifiers.

Let us consider the differences between the word encoding and the text encoding. Elements of each string vector which represents a word are text identifiers, whereas those of one which represents a text are word. The process of encoding texts involves the link of each text to a list of words, where as that of doing words does the link of each word to a list of texts. For performing semantic similarity between string vectors, in text processing, the word similarity matrix is used as the basis, while in word processing, the text similarity matrix is used. The relations between words and texts are defined as features of strings in encoding texts and words.

3.2 String Vectors

This section is concerned with the operation on string vectors and the basis for carrying out it. It consists of two subsections and assumes that a corpus is required for performing the operation. In Section 3.2.1, we describe the process of constructing the similarity matrix from a corpus. In Section 3.2.2, we define the string vector formally and characterize the operation mathematically. Therefore, this section is intended to describe the similarity matrix and the operation on string vectors.

3.2.1 Similarity Matrix

This subsection is concerned with the similarity matrix as the basis for performing the semantic operation on string vectors. Each row and column of the similarity matrix corresponds to a text in the corpus. The similarities of all possible pairs of texts are given as normalized values between zero and one. The similarity matrix which we construct from the corpus is the $N \times N$ square matrix with symmetry elements and 1's diagonal elements. In this subsection, we will describe formally the definition and characterization of the similarity matrix.

Each entry of the similarity matrix indicates a similarity between two corresponding texts. The two documents,d_i and d_j, are indexed into two sets of words, D_i and D_j. The similarity between the two texts is computed by equation (1),

$$sim(d_i, d_j) = \frac{2|D_i \cap D_j|}{|D_i| + |D_j|} \tag{1}$$

where $|D_i|$ is the cardinality of the set, D_i. The similarity is always given as a normalized value between zero and one; if two documents are exactly same to each other, the similarity becomes 1.0 as follows:

$$sim(d_i, d_j) = \frac{2|D_i \cap D_i|}{|D_i| + |D_i|} = 1.0$$

and if two documents have no shared words, $D_i \cap D_j = \emptyset$ the similarity becomes 0.0 as follows:

$$sim(d_i, d_j) = \frac{2|D_i \cap D_j|}{|D_i| + |D_j|} = 0.0$$

The more advanced schemes of computing the similarity will be considered in next research.

From the text collection, we build $N \times N$ square matrix as follows:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1d} \\ s_{21} & s_{22} & \dots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \dots & s_{dd} \end{pmatrix}.$$

N individual texts which are contained in the collection correspond to the rows and columns of the matrix. The entry, s_{ij} is computed by equation (1) as follows:

$$s_{ij} = sim(d_i, d_j)$$

The overestimation or underestimation by text lengths are prevented by the denominator in equation (1). To the number of texts, N, it costs quadratic complexity, $O(N^2)$, to build the above matrix.

Let us characterize the above similarity matrix, mathematically. Because each column and row corresponds to its same text in the diagonal positions of the matrix, the diagonal elements are always given 1.0 by equation (1).

In the off-diagonal positions of the matrix, the values are always given as normalized ones between zero and one, because of $0 \leq 2|D_i \cap D_i| \leq |D_i| + |D_j|$ from equation (1). It is proved that the similarity matrix is symmetry, as follows:

$$s_{ij} = sim(d_i, d_j) = \frac{2|D_i \cap D_j|}{|D_i| + |D_j|} = \frac{2|D_j \cap D_i|}{|D_j| + |D_i|}$$
$$= sim(d_j, d_i) = s_{ji}$$

Therefore, the matrix is characterized as the symmetry matrix which consists of the normalized values between zero and one.

The similarity matrix may be constructed automatically from a corpus. The N texts which are contained in the corpus are given as the input and each of them is indexed into a list of words. All possible pairs of texts are generated and the similarities among them are computed by equation (1). By computing them, we construct the square matrix which consists of the similarities. Once making the similarity matrix, it will be used continually as the basis for performing the operation on string vectors.

3.2.2 String Vector and Semantic Similarity

This section is concerned with the string vectors and the operation on them. A string vector consists of strings as its elements, instead of numerical values. The operation on string vectors which we define in this subsection corresponds to the cosine similarity between numerical vectors. Afterward, we characterize the operation mathematically. Therefore, in this section, we define formally the semantic similarity as the semantic operation on string vectors.

The string vector is defined as a finite ordered set of strings as follows:

$$\mathbf{str} = [str_1, str_2, ..., str_d]$$

An element in the vector, str_i indicates a text identifier which corresponds to its attribute. The number of elements of the string vector, $\mathbf{str}$ is called its dimension. In order to perform the operation on string vectors, we need to define the similarity matrix which was described in Section 3.2.1, in advance. Therefore, a string vector consists of strings, while a numerical vector does of numerical values.

We need to define the semantic operation which is called 'semantic similarity' in this research, on string vectors; it corresponds to the cosine similarity on numerical vectors. We note the two string vectors as follows:

$$\mathbf{str}_1 = [str_{11}, str_{12}, ..., str_{1d}]$$
$$\mathbf{str}_2 = [str_{21}, str_{22}, ..., str_{2d}]$$

where each element, d_{1i} and d_{21i} indicates a text identifier. The operation is defined as equation (3.2.2) as follows:

$$sim(\mathbf{str}_1, \mathbf{str}_2) = \frac{1}{d}\sum_{i=1}^{d} sim(d_{1i}, d_{2i}) \qquad (2)$$

The similarity matrix was constructed by the scheme which is described in Section 3.2.1, and the $sim(d_{1i}, d_{2i})$ is computed by looking up it in the similarity matrix. Instead of building the similarity matrix, we may compute the similarity, interactively.

The semantic similarity measure between string vectors may be characterized mathematically. The commutative law applies as follows:

$$sim(\mathbf{str}_1, \mathbf{str}_2) = \frac{1}{d}\sum_{i=1}^{d} sim(d_{1i}, d_{2i})$$

$$= \frac{1}{d}\sum_{i=1}^{k} sim(d_{2i}, d_{1i}) = sim(\mathbf{str}_2, \mathbf{str}_1)$$

If the two string vectors are exactly same, its similarity becomes 1.0 as follows:

$$\text{if } \mathbf{str}_1 = \mathbf{str}_2 \text{with } \forall_i sim(d_{1i}, d_{2i}) = 1.0$$

$$\text{then } sim(\mathbf{str}_1, \mathbf{str}_2) = \frac{1}{d}\sum_{i=1}^{d} sim(d_{1i}, d_{2i}) = \frac{d}{d} = 1.0$$

However, note that the transitive rule does not apply as follows:

$$\text{if } sim(\mathbf{str}_1, \mathbf{str}_2) = 0.0 \text{ and } sim(\mathbf{str}_2, \mathbf{str}_3) = 0.0$$

$$\text{then, not always } sim(\mathbf{str}_1, \mathbf{str}_3) = 0.0$$

We need to define the more advanced semantic operations on string vectors for modifying other machine learning algorithms. We define the update rules of weights vectors which are given as string vectors for modifying the neural networks into their string vector based versions. We develop the operations which correspond to computing mean vectors over numerical vectors, for modifying the k means algorithms. We consider the scheme of selecting representative vector among string vectors for modifying the k medoid algorithms so. We will cover the modification of other machine learning algorithms in subsequent researches.

3.3 Proposed Version of KNN

This section is concerned with the proposed KNN version as the approach to the text categorization. Raw texts are encoded into string vectors by the process which was described in Section 3.1. In this section, we attempt to the traditional KNN into the version where a string vector is given as the input data. The version is intended to improve the classification performance by avoiding problems from encoding texts into numerical vectors. Therefore, in this section, we describe the proposed KNN version in detail, together with the traditional version.

The traditional KNN version is illustrated in Figure 3. The sample words which are labeled with the positive class or the negative class are encoded into numerical vectors. The similarities of the numerical vector which represents a novice

word with those representing sample words are computed using the Euclidean distance or the cosine similarity. The k most similar sample words are selected as the k nearest neighbors and the label of the novice entity is decided by voting their labels. However, note that the traditional KNN version is very fragile in computing the similarity between very sparse numerical vectors.

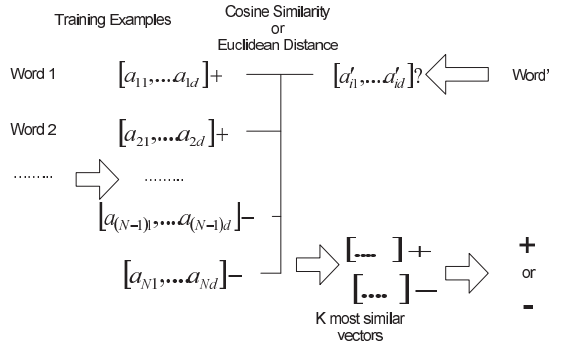


Fig. 3: The Traditional Version of KNN

Separately from the traditional one, we illustrate the classification process by the proposed version in Figure 4. The sample texts labeled with the positive or negative class are encoded into string vectors by the process described in Section 3.1. The similarity between two string vectors is computed by the scheme which was described in Section 3.2.2. Identically to the traditional version, in the proposed version, the k most similarity samples are selected, and the label of the novice one is decided by voting ones of sample entities. Because the sparse distribution in each string vector is never available inherently, the poor discriminations by sparse distribution are certainly overcome in this research.

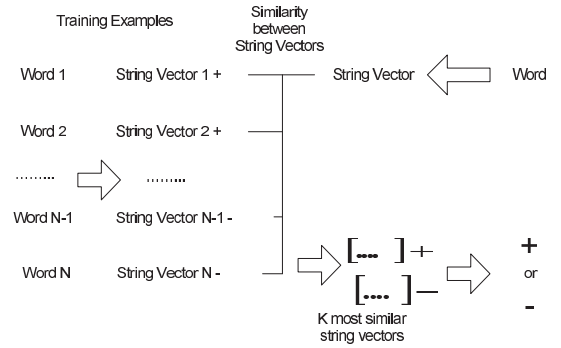


Fig. 4: The Proposed Version of KNN

We may derive some variants from the proposed KNN version. We may assign different weights to selected neighbors instead of identical ones: the highest weights to the first nearest neighbor and the lowest weight to the last one. Instead of a fixed number of nearest neighbors, we select any number of training examples within a hyper-sphere whose center is the given novice example as neighbors. The categorical scores are computed proportionally to similarities with training examples, instead of selecting nearest neighbors. We may also consider the variants where more than two variants are combined with each other.

Because string vectors are characterized more symbolically than numerical vectors, it is easy to trace results from classifying items in the proposed version. It is assumed that a novice item is classified by voting the labels of its nearest neighbors. The similarity between string vectors is computed by the scheme which is described in Section 3.2.2. We may extract the similarities of individual elements of the novice string vector with those of nearest neighbors labeled with the classified category. Therefore, the semantic similarities play role of the evidence for presenting the reasons of classifying the novice one so.

3.4 Application to Keyword Extraction

This section is concerned with the scheme of applying the proposed KNN version which was described in Section **??** to the keyword extraction task. Before doing so, we need to transform the task into one where machine learning algorithms are applicable as the flexible and adaptive models. We prepare the words which are labeled with 'keyword' or 'not' as the sample data. The words are encoded into tables by the scheme which was described in Section **??**. Therefore, in this section, we describe the process of extracting keywords from texts automatically using the proposed KNN with the view of the keyword extraction into a classification task.

In this research, the keyword extraction is viewed into a binary classification task, as shown in Figure 5. A text is given as the input, and a list of words is extracted by indexing the text. Each word is classified by the classifier into either of two labels: 'keyword' or 'not'. The words which are classified into 'keyword' are selected as the output of the keyword extraction system. For doing so, we need to collect words which are labeled with one of the two labels as sample examples, in advance.

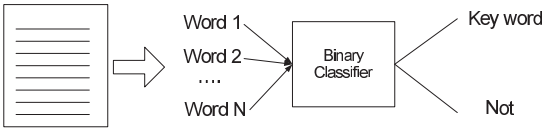


Fig. 5: View of Keyword Extraction into Binary Classification

We need to prepare sample words which are labeled with 'keyword' or 'not', before classifying a novice one or ones. A text collection is segmented into sub-collections of content based similar words which are called domains, manually or automatically. We prepare sample words which are labeled manually, domain by domain. To each domain,

we assign and train a classifier with the words in the corresponding sub-collection. When a text is given as the input, the classifier which corresponds to the most similar domain is selected among them.

We mention the process where an article is given as the input and a list of keywords is generated as the output. We nominate the classifier which corresponds to the sub-group which is similar as the given article, based on its content. A list of words is extracted by indexing the article, and each word is encoded into structured forms. The extracted words are classified by the nominated classifier into 'keyword' or 'not', and the words which are classified into the former are selected. The performance depends on the granularity of each sub-group; it should be optimized between the two factors: the amount of sample examples and the subgroup granularity.

Even if the keyword extraction is viewed into an instance of word categorization, it needs to be distinguished from the topic based word categorization. The word categorization is given as a single multiple classification or multiple binary classifications, whereas the keyword extraction is fixed only to a single binary classification. In the word categorization, each word is classified semantically into one or some of the predefined topics, whereas in the keyword extraction, it is classified into an essential word, or not. In the word categorization, each word is classified by its meaning, whereas in the keyword extraction, it is classified by its relevancy to the given text. In the word categorization, when the given task is decomposed into binary classification tasks, a classifier is assigned to each topic, whereas, in the keyword extraction, a classifier is done to each domain.

4. Conclusion

Let us mention the remaining tasks for doing the further research. The proposed approach should be validated and specialized in the specific domains: medicine, engineering and economics. Other features such as grammatical and posting features may be considered for encoding words into string vectors as well as text identifiers. Other machine learning algorithms as well as the KNN may be modified into their string vector based versions. By adopting the proposed version of the KNN, we may implement the keyword extraction system as a real program.

5. Acknowledgement

This work was supported by 2016 Hongik University Research Fund.

References

[1] F. Sebastiani, "Machine Learning in Automated Text Categorization", pp1-47, ACM Computing Survey, Vol 34, No 1, 2002.
[2] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification with String Kernels", pp419-444, Journal of Machine Learning Research, Vol 2, No 2, 2002.
[3] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch String Kernels for Discriminative Protein Classification", pp467-476, Bioinformatics, Vol 20, No 4, 2004.
[4] R. J. Kate and R. J. Mooney, "Using String Kernels for Learning Semantic Parsers", pp913-920, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006.
[5] T. Jo and D. Cho, "Index based Approach for Text Categorization",International Journal of Mathematics and Computers in Simulation, Vol 2, No 1, 2008.
[6] T. Jo, "Single Pass Algorithm for Text Clustering by Encoding Documents into Tables", pp1749-1757, Journal of Korea Multimedia Society, Vol 11, No 12, 2008.
[7] T. Jo, "Device and Method for Categorizing Electronic Document Automatically", Patent Document, 10-2009-0041272, 10-1071495, 2011.
[8] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", pp839-849, Soft Computing, Vol 19, No 4, 2015.
[9] T. Jo, "Inverted Index based Modified Version of K-Means Algorithm for Text Clustering", pp67-76, Journal of Information Processing Systems, Vol 4, No 2, 2008.
[10] T. Jo, "Representationof Texts into String Vectors for Text Categorization", pp110-127, Journal of Computing Science and Engineering, Vol 4, No 2, 2010.
[11] T. Jo, "NTSO (Neural Text Self Organizer): A New Neural Network for Text Clustering", pp31-43, Journal of Network Technology, Vol 1, No 1, 2010.
[12] T. Jo, "NTC (Neural Text Categorizer): Neural Network for Text Categorization", pp83-96, International Journal of Information Studies, Vol 2, No 2, 2010.

Performance of Compression Algorithms used in Data Management Software

Pierre Andre Bouabre
Department of Informatics
University of South Carolina Upstate
800 University Way, Spartanburg, SC 29303
bouabre@email.uscupstate.edu

Tyrone S. Toland
Department of Informatics
University of South Carolina Upstate
800 University Way, Spartanburg, SC 29303
ttoland@uscupstate.edu

Abstract – Information growth is expanding to an almost unmanageable state in today's society. A challenging and active area of information management is data compression. This paper provides an analysis of compression and decompression algorithms. In particular, this research analyzes the run time efficiency and performance of compression algorithms. Several commercial compression algorithms (e.g., Huffman, Flate/Deflate, LZW) were used to gather empirical results. The results in this study show that the algorithms performance (i.e., run time, compression efficiency) can be different when executed either individually or combined.

Keywords— Compression algorithms, lossy compression, lossless compression, Huffman , Flate/Deflate, LZW compression

I. INTRODUCTION

Technology has given society the ability to learn and adapt to the environment. Collecting data helps to provide answers to challenging questions; moreover, collecting and managing data is essential to the sustainability of society. Data collected over many years poses the following questions for society: "How can data storage methods be improved for easy accessibility?"; "How much space can be saved when storing information?"; and "How much time does it take to process and access data?"

The exponential growth of the amount of data is rising quickly. Due to the yearly rate of increase managing data can be difficult. The large volume of data growth poses both a technological and economical dilemma, especially in relation to storage space and processing time. In fact, large amounts of data collected over years can make an organization's data storage management expensive. Equation 1 shows the computed cost of outsourcing data storage [1, 4].

$$TotalCost = InitialCost + FloorCost + EnergyCost + ServiceCost + DisposalCost + EnvironmentalCost \quad (1)$$

Researchers have developed many ways to improve data storage processing. In [8], methods such as 1) Direct Attached Storage (DAS) (subsystems to store data are linked locally on a computer), 2) Network Attached Storage (NAS) (subsystems to store data linked to a network via simple file-serving appliance), and 3) Storage Area Networks (SAN) (subsystems to store data linked together on a network) are used to access other networks for storage. While these systems solve some of the data storage space issue, storage space can still be expensive in the long run.

Reducing the amount of space by compressing data can help reduce the storage cost and save some storage space. Data compression can bring an array of benefits if performed effectively [3]. Accessing compressed data relies on the amount of time it takes to decompress the data; also, the compression time for a larger data set can take additional processing time. The problem with saving space with data compression raises the competing issues of processing time vs. compression efficiency.

To be able to compress and decompress as fast as possible without losing data, the best algorithm must be selected. This research provides an analysis of compression and decompression (compression) algorithms. In particular, this research analyzes the run time efficiency and effectiveness of compression algorithms. Several commercial compression algorithms (e.g., Huffman, Flate/Deflate, LZW) are used in this study.

This paper is organized as follows. In Section II, related work is discussed. In Section III, an overview of the compression and decompression process is presented. Section IV discusses the analysis. Section V concludes the paper.

II. RELATED WORK

This section discusses research using compression algorithm methods for data storage. The more efficient data compression methods used in data storage eliminates the redundancies of data items in order to improve storage space and to reduce data storage cost [20]. Lossless data compression algorithms are considered the best approach to encode and decode data without losing data in the process.

File system compression is a lossless algorithm that compresses every data component [20]. This algorithm originates from DiskDoubler and SuperStor methods; these methods were used in early computers to support hard drives that had limited storage capacity. The disadvantage of file system compression is that processing running time is high [2, 20].

NetApp and Rival EMC Corp are companies that propose data compression solutions [17]. NetApp proposes technology to address the increasing demand of managing data storage. The technology proposed by NetApp is considered as one of the most efficient data storage system; however, the technology does not perform well when 1) locating data (i.e., files or directories) on tape or 2) restoring data from tape [17]. To compete with NetApp, Rival EMC Corp developed an application called Celerra Data Deduplication, which focuses on data duplication and data storage; this application compresses data before handling deduplication of data [17].

Strom and Wennersten [19] discuss that lossless compression of already compressed textures, due to the fact that texture codecs are usually not adept to pausing, involves an issue texture which is downloaded over a network or reading on a disc. Texture compression aids rendering by minimizing the footmark in graphics memory. The solution proposes to address compression and decompression efficiency is to predict compression parameters [19]. The limitation encountered in the proposed solution was that the system could only resolve the slow transmission time of data over a network but could not improve the graphics memory footprint [19].

Peel, Wirth, and Zobel [11] discuss how their scheme accomplishes a more efficient (i.e., better scale) compression for larger data than current compression systems. The space necessary for their compression algorithm is sub-linear to the data size. They accomplish amelioration by achieving compression of multiple files, several times better than the compression of gzip or 7-zip. Although their application compresses data faster than 7-zip, Peel et al. [11] only compared their application against a small number of systems. Also, the proposed application in [11] avoids reading other files while compressing new data.

Millard, Nunez and Mulvane [9] discuss a hardware solution, which is a high- performance application for a parallel multi-compressor chip. This research shows that input and output choices can have a negative effect on performances regarding the routing strategies, which shows that the scheme of parallel compression system is affected by the compression performance system [9]. To solve this issue, Millard et al. [9] proposes a scalable compression solution to be used at throughputs into the field programming gate array hardware cable to handle the modern high-bandwidth system for intensive data processing. This process may prove successful, but the performance and run time are still in question.

Throughout the research done on data compressing and decompression, all the solutions have running time issues between compressing and decompressing information. Fewer solutions have issues maintaining information integrity when restoring compressed data to its original state that can be handled with a Lossless compression algorithm.

III. COMPRESSION AND DECOMPRESSION OVERVIEW

Data compression is the series of methods to encode data into fewer bits; data decompression is a process of restoring encoded data back into its initial state. See Fig. 1.

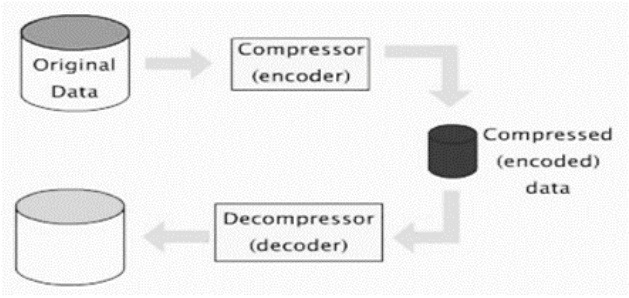


Fig. 1. Data compression and decompression example taken from [5]

Data compression presses the data to allow a smaller 1) amount of disk space in the storage unit and 2) bandwidth on the data broadcast channel. The best examples of systems using data compression are network routers, phone and most electronic systems of information exchange. Data compression is valuable because it allows faster transference of data than uncompressed data. Because compressed data takes less space, it is also cost efficient to store. In data compression there are two eminent compression concepts in use which are lossy compression and lossless compression [5, 6, 10].

Lossy compression is the class of data encoding that uses a limited quantity of data discarding procedures to symbolize the data content. Using the lossy method to compress and decompress data may affect data integrity by converting the data into a slightly different state than the original state, i.e., decompressed data may not be a completely restored. However, this slightly different state is sufficient to use in the compression process; all the bits of data remain in the data file after the data is decompressed which guarantees that the data is not lost. The lossy compression algorithm removes information that is considered insignificant from the original data state when it performs the compression process. The algorithm builds the data by using space efficiently to produce an efficient data format; it also generates a much smaller compressed data file than the lossless method [5, 6, 10].

Lossless compression is a data compression method that permits the original compressed data to be decompressed without losing data integrity. Lossless data compression algorithms find and repeat patterns to ultimately reduce data redundancy before encoding the data. If data redundancy is high in the input data, then the file size of the compressed data will be low [5].

IV. ANALYSIS

A. Experiment Overview

This research shows the problem of saving cost and space with data storage using the best data compression algorithm available. Compression and decompression with data storage accessibility (CDWDSA) are concepts used by NetApp, Rival EMC, and other modern systems implementing data compression in their framework. Although solutions are proposed by modern systems, they still are unfixed answers to the question: "What is the best algorithm to use in a CDWDSA environment?" This study believes that algorithm performance needs to 1) look at data compression needs with

data storage and 2) data decompression with data accessibility. Data storage and accessibility analysis is based on how fast data can be stored and accessed using compression algorithms without affecting the data integrity. This paper proposes to test different compression Lossless algorithms to analyze their run time performances and similarity. This study focuses on Lossless algorithms because these algorithms ensure data integrity.

This analyses will compare the performance of the following Lossless algorithms:

- The Huffman algorithm encodes the original data using codes containing the length of the duplicate data patterns. This algorithm uses a tree data structure to encode and decode data (e.g., compression, decompression) [16].
- The LZW compression algorithm uses the LZ77 algorithm method to compress and decompress data using a compression dictionary. The LZ77 uses a dictionary based compression process that utilizes pointers to identify duplicate components in a data set. These pointers are to compress and decompress data [7, 14, 18].
- The Flate/Deflate algorithm combines the Huffman algorithm and the LZ77 algorithms to form a more versatile and intelligent compression and decompression process [7, 14, 18].

To run this experiment, several programs where developed to implement each algorithm (i.e., Flate/Deflate, Huffman, LZW). The programming language that was used was C#. Each program recorded the performance of the algorithms. The performance was measured by the amount of time in milliseconds that each algorithm took to execute the compression and decompression process on different file sizes.

B. Discussion

Table I shows that the LZW compression algorithm has reduced the total amount of data used in this experiment by 22.361% which is slightly better than Flate/Deflate and Huffman. That is, LZW was 0.002 % better than Flate/Deflate and 15.771% better than Huffman.

TABLE I. COMPARING COMPRESSION SIZE

Algorithm	Total of 200 files (bytes)	Compressed data (bytes)	Space saved (bytes)	Space saved
Huffman	1303106371	1217188619	85917752	6.59%
Flate/ Deflate	1303106371	1011689640	291416731	22.363%
LZW	1303106371	1011712440	291393931	22.361%

Table II compression results show that Huffman ran 10% faster than LZW and 39% faster than Flate/Deflate. Table III decompression results show that Huffman also ran 10% faster than LZW and 39% faster than Flate/Deflate.

TABLE II. COMPARING COMPRESSION RUNNING TIME

Algorithm	Total of 200 files (bytes)	Compressed data time (millisecond)	Ranking
Huffman	1303106371	249338	1
Flate/ Deflate	1303106371	345401	3
LZW	1303106371	276028	2

TABLE III. COMPARING DECOMPRESSION RUNNING TIME

Algorithm	Total of 200 files (bytes)	Compressed data time (millisecond)	Ranking
Huffman	1303106371	249338	1
Flate/ Deflate	1303106371	345401	3
LZW	1303106371	276028	2

Tables II and III show that each algorithm takes approximately the same amount of time to decompress and compress the data files. That is, Huffman takes 249338 milliseconds to compress and decompress, LZW takes 276028 milliseconds to compress and decompress, and Flate/Deflate takes 249338 milliseconds to compress and decompress.

Flate/Deflate, while consisting of both LZW and Huffman, produced a compressed file size that is similar to LZW; however, the Flate/Deflate algorithm running time was higher than either LZW or Huffman.

This analysis suggests that if execution speed is a factor, then the Huffman algorithm should be used; if compression effectiveness (i.e., smallest compressed file size) is a factor, then either LZW or Flate/Deflate can be used.

V. CONCLUSION

This paper presented an analysis of Lossless data compression algorithm. In particular, this paper compared the Huffman, Flate/Deflate, and LZW algorithms. This paper compared the running time and the effectiveness of the algorithms.

Future work could include running experiments that implement the compression algorithms in different programming languages to study the run time and compression performance. Additional future work should also include analyzing additional Lossless algorithms. A larger set of test cases should provide additional information concerning the robustness of the algorithms.

REFERENCES

[1] Advantages and disadvantages of tape backup. (2013, May 1). Retrieved from 2015 https://library.netapp.com/ecmdocs/ECMP1196992/html/GUID-79D26031-7D05-4E1B-B7A7-3FBE8089B3A0.html.

[2] Barett, J. (n.d.). What Are the Advantages & Disadvantages of Using File Compression? Retrieved September 17, 2015, from http://smallbusiness.chron.com/advantages-disadvantages-using-file-compression-27740.html.

[3] Brinkmann, B., Bower, M., Stengel, K., Worrell, G., & Stead, M. (n.d.). Large-scale electrophysiology: Acquisition, compression, encryption,

and storage of big data. Journal of Neuroscience Methods, Vol. 180 (1), 185-192.

[4] Dutta, A., & Hasan, R. (2014). How Much Does Storage Really Cost? Towards a Full Cost Accounting Model for Data Storage. Economics of Grids, Clouds, Systems, and Services Lecture Notes in Computer Science, 29-43.

[5] Compression Concepts. (2010). Retrieved October 2, 2015, from http://www.gitta.info/DataCompress/en/html/CompIntro_learningObject2.html.

[6] How Does File Compression Work. (2015). Retrieved October 2, 2015, from http://www.makeuseof.com/tag/how-does-file-compression-work/.

[7] Huffman, D. (1952). A Method for the Construction of Minimum-Redundancy Codes. Proceedings of the IRE Proc. IRE, 40(9), 1098-1101. Retrieved September, 2015.

[8] JAKUB, S. (2011). ENTERPRISE DATA STORAGE: A REVIEW OF SOLUTIONS. (Issue 41), 222-235. Retrieved from Academic Search Complete.

[9] Milward, M., Nunez, J., & Mulvaney, D. (2004). Design and implementation of a lossless parallel high-speed data compression system. IEEE Trans. Parallel Distrib. Syst. IEEE Transactions on Parallel and Distributed Systems, 481-490.

[10] Ladino, J. (1996). Data Compression Algorithms. Retrieved December 4, 2015, from http://www.ccs.neu.edu/home/jnl22/oldsite/cshonor/jeff.html.

[11] Peel, A., Wirth, A., & Zobel, J. (2011). Collection-based compression using discovered long matching strings. Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11, Pages 2361-2364.

[12] Peterson, B. (2008, July 1). Top five data storage compression methods. Retrieved from http://searchitchannel.techtarget.com/feature/Top-five-data-storage-compression-methods.

[13] Rouse, M. (n.d.). Lossless and lossy compression. Retrieved December 4, 2015, from http://whatis.techtarget.com/definition/lossless-and-lossy-compression.

[14] S. B. (2002, March 19). LZW Data Compression. Retrieved November 06, 2015, from https://www.cs.duke.edu/csed/curious/compression/lzw.html.

[15] Service-Oriented Architecture (SOA) Definition. (2000). Retrieved September 3, 2015, from http://www.service-architecture.com/articles/web-services/service-oriented_architecture_soa_definition.html.

[16] Shahbahrami, A., Bahrampour, R., Rostami, M. S., & Ayoubi, M. (2011). Evaluation of Huffman and Arithmetic Algorithms for Multimedia Compression Standards. IJCSEA International Journal of Computer Science, Engineering and Applications, 1(4), 34-47. Retrieved December 03, 2015, from http://arxiv.org/ftp/arxiv/papers/1109/1109.0216.pdf. ITAL Information Technology and Libraries, Vol. 28(Issue 3), P143-153. Retrieved November 30, 2015, from Education Full Text (H.W. Wilson).

[17] Sliwa, C. (2009, December 1). Primary storage data reduction advancing via data deduplication, compression. Retrieved September 16, 2015, from http://searchstorage.techtarget.com/report/Primary-storage-data-reduction-advancing-via-data-deduplication-compression.

[18] Suthers, D. (2014, March). Greedy Algorithms. Retrieved December, 2015, from http://www2.hawaii.edu/~nodari/teaching/s15/Notes/Topic-13.html.

[19] Strom, J., & Wennersten, P. (2011). Lossless compression of already compressed textures. Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics - HPG '11, Pages 177-182.

[20] Top five data storage compression methods. (2008, July 1). Retrieved October 9, 2015, from http://searchitchannel.techtarget.com/feature/Top-five-data-storage-compression-methods.

[21] Types of Storage. (2012, July 17). Retrieved from http://typesofbackup.com/types-of-storage/.

BlueInsight: A Community & Information-Centric Web Platform for Benchmarking Community Shared Data

En Cheng
Department of Computer Science
University of Akron, Akron, OH 44325

Anshul Sheopuri
Thomas J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

BlueInsight is a Community & Information-Centric Web platform for benchmarking community shared data. BlueInsight has a three-layered architecture, which consists of database independent data access services, scalable data analytics services, and delegation-enabled reporting services. We also present (i) a sampling method that trades off the time cost of data extraction and analysis and statistical confidence to provide scalable services for benchmarking and (ii) correlated analytics as an enhancement to the core capability of BlueInsight.

Keywords

Benchmarking, Community, Scalability, Correlated Analytics

1. INTRODUCTION

Community & Information-Centric (CIC) Web platforms have gained importance as they provide enterprises the ability to use the Web as a medium to collaborate, share data as well as services amongst a community with common interests. This motivates us to study the problem of developing new paradigms for enterprises to make better use of the shared data. We present BlueInsight, a CIC Web platform for benchmarking community shared data. BlueInsight is useful in a wide range of application areas, including Travel & Entertainment (T&E) expense Management, Human Resource Management and Lending Services. An emerging application of BlueInsight is benchmarking business controls, policies and metrics of interest, where the shared data is analyzed to derive actionable insights for the benefit of the contributing organizations. For example, a company in the Lending Services domain can benefit from benchmarking the interest rates it provides to low net-worth customers against an appropriate community of companies, as part of an effort to evaluate its risk from sub-prime loans.

In this paper, we demonstrate the value of BlueInsight through an illustrative implementation in the context of T&E expenses. Employees of most medium to large size companies submit expense reports corresponding to corporate expenses that they incur such as food, travel, and lodging. These expenses are expected to be in line with certain pre-defined guidelines or business rules established by the companies, and may be subject to further review by auditors. In addition to auditing expense reports as part of compliance and controls, a company might be interested in leveraging community shared data to evaluate its business rules and vendor contracts. For example, by determining that the company is incurring excessive hotel expenses in a particular geography compared to its community, the company may use this information to renegotiate vendor contracts for that geography. In this context, we define the community of companies to comprise those that maintain and are willing to share employee travel expense data.

Our contributions are summarized as follows:

- We introduce database independent data access services, scalable data analytics services, and delegation-enabled reporting services for BlueInsight.
- We develop a framework for sampling client data to optimize performance and confidence in our estimates considering the number of companies in the community.
- We introduce correlated analytics to strengthen the core benchmarking functionality of BlueInsight.

2. The BlueInsight System

We implemented a prototype of BlueInsight in the context of T&E expenses to demonstrate the business value of the CIC platform. Below, we first describe the architecture of our prototype. We then discuss the end-to-end process flow from the perspective of a member of the community of companies.

2.1 Architecture

Our prototype has a three-layered architecture illustrated in Fig. 1, which we discuss in detail below.

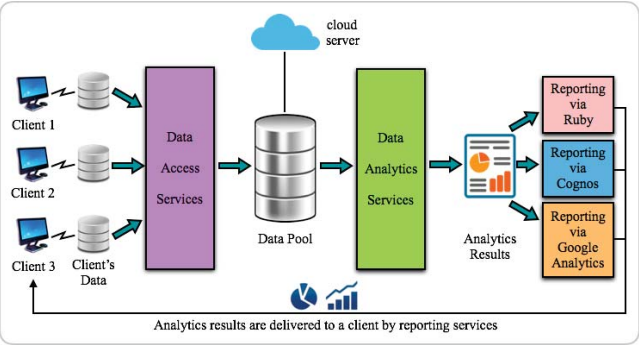


Fig. 1. BlueInsight Architecture

Data Access Services (DAS1): DAS1 performs database access, which extracts clients' data from their local database to the global scale database hosted by BlueInsight. A desirable property of DAS1 in BlueInsight is database independence, which improves the flexibility of the platform and allows DAS1 to extract data from MySQL, DB2, Sqlite, etc. BlueInsight addresses this problem with ActiveRecord, a standalone object-relational mapping package for Ruby which is part of the web-application framework Rails. ActiveRecord performs database access using objects. Ruby on Rails provides BlueInsight insulation from the underlying database through object-relational mapping, i.e., mapping objects to a relational database. In addition, BlueInsight can benefit from data distribution in Internet scale data centers.

DASI can effectively access data distributed around the globe for aggregate analysis.

There are two distinct approaches to extracting client data: (1) Extracting the entire client T&E data, and, (2) extracting a subset of the data or a statistic depending on the benchmarking scenario. We refer to the first approach as *r*-DAS1 and the second as *m*-DAS1. The benefit of *m*-DAS1 is that data confidentiality risks are reduced but the con is that every time a new metric is developed or an existing metric is enhanced, potentially a different subset of data needs to be extracted.. Further, some metrics may not even have sufficient statistics with a reasonable state space. The current implantation of BlueInsight follows *r*-DAS1. However, it is easy to adopt *m*-DAS1 with simple modifications.

Data Analytics Services (DAS2): DAS2 contains statistical algorithms that are determined by the business scenarios of interest. A desirable property of DAS2 in BlueInsight is scalability. Inspired by the idea of leveraging a "cloud" to meet business requirements, BlueInsight can benefit from cloud computing, e.g. Amazon's Elastic Compute Cloud (EC2) [1], and Google Cloud Platform [2], in order to provide scalable data analytics.

In terms of exploiting Amazon EC2, one practical approach for BlueInsight is to choosing Vertica Analytic Database [8] as CIC's back-end, as it is a Cloud-based, Grid-enabled columnar analytic database hosted on Amazon's Elastic Compute Cloud. The Vertica Analytic Database's scalability, flexibility and ease of use can be demonstrated by its wide range of customers like JP Morgan Chase, Verizon, Mozilla, Comcast, Level 3 Communications and Vonage. Another promising solution to the scalability challenge is to employ sampling. This is discussed in detail in section 3.1.

Reporting Services (RS): RS delivers the analytics results of DAS2 to clients in a user-friendly interface. Ruby on Rails provides BlueInsight with several graphical reports, which are implemented in BlueInsight. BlueInsight is also capable of delegating reporting responsibilities to existing RS, including Cognos [3] and Google Analytics [4]. By integrating thirty-party RS, the results of DAS2 may be reported by any suitable RS. Thus, the capability of integrating thirty-party RS enlarges the applicability of BlueInsight since the architecture is agnostic to the specific RS that we employ.

2.2 Process Flow

We now provide an illustration of the prototype developed with the standard process flow summarized in Fig. 2. At the login interface, a representative of a given company enters his information for access; if both username and password are correct, the client will be redirected to the interface for sharing data. The client is first prompted to enter information pertaining to where the data resides. Then, the client enters information of the benchmarking scenario of interest (e.g. expense type and location) and the algorithm to enable benchmarking. Finally, the client may view the following: (1) The results of the algorithm, for example, finding that the client's data is an outlier with respect to the collective data from the community of companies for that scenario, and (2) A visual representation of the client data and the community-shared data for that scenario, for example, histograms, to enable visual confirmation of the result in (1). Further, other scenarios may be recommended for evaluation depending on the result in (1) (see Section 3.2. for details).

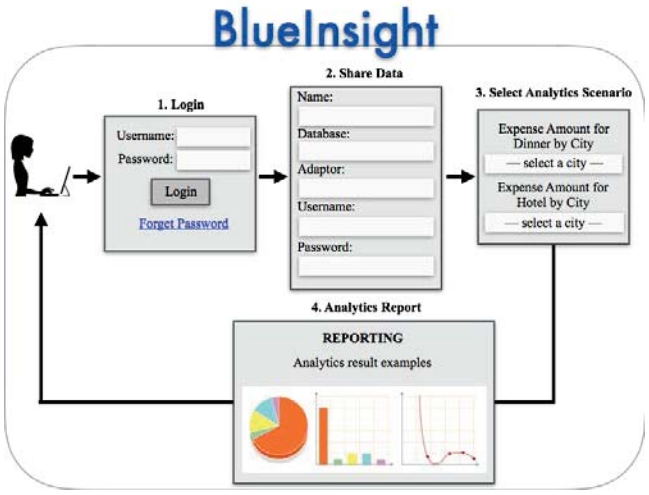


Fig. 2. Process Flow for BlueInsight

3. Demonstration Features

In this section, we discuss two important features of our prototype. First, we address the issue of scalability through a framework for sampling that we have developed. Second, we introduce correlated analytics as an enhancement to the core capability of BlueInsight.

3.1 Scalability

Given the large volume of data employed for robust benchmarking today, scalability is a desirable property of BlueInsight. Given a benchmarking scenario, a large volume of T&E expense data results in increased time cost, as well as increased confidence in statistical estimates of our benchmarking result. In addition to leveraging cloud computing in BlueInsight, data analytics services can also benefit from sampling by means of a tradeoff between time cost of data extraction and analysis and statistical confidence. There are several approaches to determining the appropriate sample size (see, for example, [5] and [6]). The framework of these approaches differs from BlueInsight since they do not incorporate the number of clients as an input for sampling.

We outline our solution approach to the scalability problem with an illustrative scenario in the context of dinner expenses. Assume that there are n clients, R_1, R_2,…, R_n in the community. Our objective is to determine the number of records to sample for a new client R_{n+1} in order to optimize the time cost of a statistical test, statistical confidence of the result, an increasing function of the number of clients and the number of data points that a client contributes.

Given a statistical test, for example, the chi-square test, the time cost for the client R_{n+1} may be written as $t = f(x_{n+1})$, where $f(.)$ is an increasing function and x_i is the number of records contributed by client i, i = 1,2,…, n+1. For the chi-squared test (for histogram matching), the time cost for chi-square test may be

approximated by a linear function, $t = k_1 x_{n+1} + c$, where k_1 and c are non-negative constants (see Fig. 3).

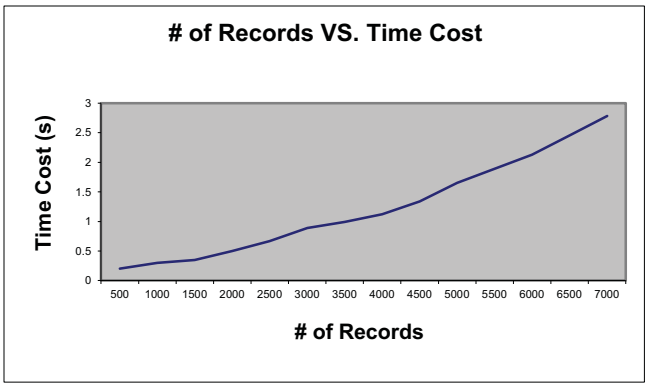


Fig. 3. An illustrative instance of time cost for chi-square test

Suppose that the client desires a 95 % confidence in estimates of his results for benchmarking using the chi-squared test. Let T_1 and T_2 be the minimum number of records from client R_{n+1} and the number of clients in the community respectively that would result in the desired confidence.

Now, we can evaluate x_{n+1} by minimizing

$$f(x_{n+1}) + g((T_1 - x)^+, (T_2 - \sum_{i=1}^{n} I\{x_i \geq x\})^+),$$

where g(.) is bi-variate penalty function of the number of records less T_1 of client R_{n+1} and the number of clients less T_2 in the community, $I(.)$ is the indicator function and

$$x^+ = \begin{cases} x & x \geq 0; \\ 0 & x < 0. \end{cases}$$

The details on estimation of $g(.)$, T_1 and T_2 are omitted due to lack of space.

3.2 Correlated Analytics

In this section, we describe an enhancement to the core capability of BlueInsight with correlated analytics where the shared data can be exploited to identify targeted insights i.e., to provide suggestions or recommendations to a participating client of new scenarios to benchmark. For example, if the dataset contains categorical data, and a particular category of data for a company is identified, through a statistical test, to exhibit abnormal behavior with respect to the community, we can learn from the shared data what other data types for this company have a greater likelihood to also exhibit abnormal behavior.

Our focus is the application of a recommendation system in the context of a CIC platform for Business-to-Business (B2B) applications. While recommendation systems have been widely applied to provide suggestions to users for new items to buy, employing recommendations about what else needs to be investigated for an outlier behavior has not been studied, especially in the CIC context. Our approach is based on estimating the joint probability distributions of Bernoulli random variables that denote whether a category of an attribute for a company is an outlier or not with respect to the community. While making recommendations to a company to identify an outlier in another category of the attribute, we leverage these estimated distributions in the spirit of classical Collaborative Filtering methods [7].

In this paper, we only provide an intuitive explanation of our correlated analytics approach with an illustrative scenario in the context of dinner expenses. The first step is to determine, for each company, whether its expenses for each geographical location are an outlier or not. For example, consider the expenses corresponding to the New York geography of each company in the community. The idea of this step is as follows. For the company, we first estimate the empirical distribution of its dinner expenses for New York. We next estimate the distribution of dinner expenses of the community of companies for New York. We then ask whether the distribution of dinner expenses for the geography of interest, New York, is the same for the company and the pool to determine whether the company is indeed an outlier or not. Next, we estimate the joint probability distributions of outlier behavior of dinner expenses of the various geographies for a company. Recommendations of correlated analytics are provided based on conditional probability distributions that are derived from the joint probability distributions. .

4. REFERENCES

[1] Amazon Elastic Compute Cloud, Virtual Grid Computing. http://www.amazon.com/gp/browse.html?node=201590011

[2] Google Cloud Platform. https://cloud.google.com/

[3] Cognos, http://www.cognos.com/

[4] Google Analytics, https://www.google.com/analytics/

[5] Cochran, W. G. 1963. Sampling Techniques, 2nd Ed., New York: John Wiley and Sons, Inc.

[6] Bartlett, J. E., II, Kotrlik, J. W., & Higgins, C. (2001). Organizational research: Determining appropriate sample size for survey research. *Information Technology, Learning, and Performance Journal*, 19(1) 43-50.

[7] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3), pp. 77-87.

[8] Vertica Analytic Database, http://www8.hp.com/us/en/software-solutions/advanced-sql-big-data-analytics/

ISBN: 1-60132-441-3, CSREA Press ©

SESSION

EDUCATIONAL STRATEGIES AND E-LEARNING

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

Information retrieval at the PPGEGC Theses and Dissertations Library Catalog

Guilherme Bertoni Machado[1], Larissa Mariany Freiberger Pereira[1], José Leomar Todesco[1], Alexandre Leopoldo Gonçalves[1]
[1]Graduate Program in Engineering and Knowledge Management, Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil

Abstract - *This paper proposes a hierarchical navigational taxonomy structure to organize the documents available at the Theses and Dissertations Library Catalog of the Graduate Program in Engineering and Knowledge Management (PPGEGC) at the Federal University of Santa Catarina. Currently the PPGEGC Library Catalog does not have a hierarchical structure to organize these documents, making the information retrieval difficult. The taxonomic structure that we propose was developed based on the concentration areas (Knowledge Engineering, Knowledge Management and Media and Knowledge), as well as nine research lines that compose PPGEGC. The terms that compose the proposed taxonomy were collected based on data mining technique using the R software and then we used the keywords in the documents to complement the taxonomic structure. After all, we could organize all topics covered by dissertations and thesis in PPGEGC in a taxonomy composed by 583 terms.*

Keywords: Information Retrieval; Taxonomy; Controlled Vocabulary; Library Catalog.

1 Introduction

The volume of information produced and made available in the Web has gradually grown and, in this context, [1] conceptualizes a new phenomenon in the information age, the "Data Smog" or "data pollution" as an exaggerated amount of information available to users. Because that volume of information is not organized, users end up getting confused and stressed and the seemingly simple task of recovering information turns out to be exhaustive.

In light of this situation, nowadays techniques and tools as Controlled Vocabularies have been developed to assist in the process of information retrieval [2]. Among them we can mention Taxonomies, Ontologies, Thesaurus, and others.

According to [3], p. 28, "An alternative to classical search engines (keyword-based search) is to allow the user to navigate through contents. This process of navigation requires the user to know the way in which the information is organized. A practical solution to this inconvenient is the use of a pre-established taxonomy".

In this article, we propose a type of controlled vocabulary, named taxonomy, in order to organize the documents available at the PPGEGC Theses and Dissertations Library Catalog that, currently, does not have a hierarchical structure to organize these documents and make the information retrieval easier.

2 The PPGEGC

The Graduate Program in Engineering and Knowledge Management (PPGEGC) was created by the Federal University of Santa Catarina in May 2004. One of the main goals of PPGEGC department is to implement models, methods and techniques to promote development in organizations public and private, as in general society, trough code, manage and disseminate knowledge (explicit and tacit).

According the program mindset, knowledge is perceived as a product, process and result of social and technological interactions between human and technological agents [4].

It is an interdisciplinary program covering three main concentration areas, which were "Knowledge Management", "Knowledge Engineering" and "Media and Knowledge". Each of these areas is composed by three research lines.

The concentration area of Knowledge Management consists of the following research lines: "Theory and Practice in Knowledge Management", "Knowledge Management, Entrepreneurship and Technological Innovation" and "Sustainability Knowledge Management".

Whereas the concentration area of Knowledge Engineering consists of the following research lines: "Theory and Practice in Knowledge Engineering", "Knowledge Engineering Applied to Organizations" and "Knowledge Engineering applied to Electronic Government".

Finally, the concentration area of Media and Knowledge consists of the following research lines: "Theory and Practice

in Media and Knowledge", "Media and Dissemination of Knowledge" and "Media and Knowledge in Education".

The PPGEGC made available in electronic format all defended and approved theses and dissertations of the program in their page named "Banco de Teses e Dissertações do EGC" [5]. This database had, until August 14, 2015, 322 documents (153 theses and 169 dissertations) ready to be retrieved.

These documents are linked to two subjects: (1) thesis/dissertation and (2) concentration area. They are also indexed through tags. These tags correspond to the keywords of the document. The main difficulty in the recovery task of these documents is the fact that there is no hierarchy of concepts or keywords.

In other words, there is no structure able to organize these documents. Another problem is that different keywords can refer to the same concept, for example, "e-Gov" and "Electronic Government" correspond to the same subject and the current form of indexing database and search system are not able to understand this.

Consequently, we propose an index structure that organizes documents in PPGEGC Theses and Dissertations Library Catalog, facilitating the retrieval of these documents.

3 Taxonomy

According to [6], the term "taxonomy" began to be used in the eighteenth century by Carl von Linne to describe a hierarchical classification system for life forms. Currently this term is used to designate a controlled vocabulary whose terms are organized in a hierarchical structure, in order to organize them and make simple information retrieval in the area of knowledge mapped by taxonomy.

As reported by [7], taxonomies can be constructed from three distinct structures:

1. Descriptive Taxonomy;
2. Navigational Taxonomy;
3. Data Management Vocabulary.

The descriptive taxonomy supports the information retrieval through a basic set of controlled vocabularies. Thus, the entire contents of a particular domain of knowledge can be described through selected metadata from those authorized vocabularies.

The navigational taxonomy consists of a hierarchical structure that enables the information discovery during navigation in its own structure and it is widely used to assist final users of taxonomy in the specific knowledge domain to find the information they need.

Finally, the data management vocabulary consists of a limited list of authorized terms that are not arranged in a hierarchical structure, according to [7]. For that reason, we used the navigational taxonomy structure to construct the proposed index structure.

4 Methodology

The construction of this taxonomy occurred in four distinct stages (Fig. 1), namely:

1. Planning;
2. Vocabulary Survey;
3. Organization of Concepts;
4. Final Presentation.

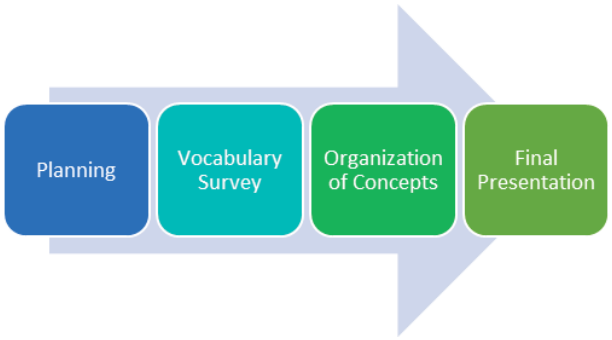


Fig. 1. Methodology Used For The Construction Of The Proposed Taxonomy

In the first stage we delimited the area of operation of this taxonomy. Therefore, we defined it was a taxonomy in the context of the PPGEGC Theses and Dissertations Library Catalog.

We also define the starting point for the construction of taxonomy, selecting basic and fundamental terms that rise to other terms collected.

In this sense, we defined that the taxonomic structure would be built from three major research areas of PPGEGC: Knowledge Management, Knowledge Engineering and Media and Knowledge and thus organize the research lines mentioned above within these large areas.

1. Knowledge Management
   a. Theory and Practice in Knowledge Management
   b. Knowledge Management, Entrepreneurship and Technological Innovation
   c. Sustainability Knowledge Management

2. Knowledge Engineering
    a. Theory and Practice in Knowledge Engineering
    b. Knowledge Engineering applied to Organizations
    c. Knowledge Engineering applied to Electronic Government
3. Media and Knowledge
    a. Theory and Practice in Media and Knowledge
    b. Media and Dissemination of Knowledge
    c. Media and Knowledge in Education

In the second stage, we survey the terms in all documents that were available at the PPGEGC Theses and Dissertations Library Catalog trough text mining. As said by [8], in p. 01, "text mining encompasses a vast field of theoretical approaches and methods with one thing in common: text as input information".

According to [9], p. 01, "text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools".

The text mining process involves techniques as ontology creation, text classification, taxonomies creation, etc and it is quite important to resolve information retrieval problems, according to [8].

The process was as follows: the documents were downloaded from the repository of PPGEGC Theses and Dissertations Library Catalog in .pdf format and later converted to .txt format.

After this process, the documents were organized in documents folders according to the three areas of knowledge and nine research lines previously described. Then we used the R software to mining data in the documents, listing the 20 most frequent words in each one of these.

In the third stage we worked in the organization of terms. At this stage we turn our efforts mainly to identify different terms that had the same meaning (synonyms) and the actual construction of navigational taxonomic structure from the terms identified in the previous step. This process is better described in the results session.

Finally, we develop the final document to present taxonomy. The model chosen was a conceptual map which shows a structured list of all the navigational taxonomy, illustrated by appendix A, B and C.

5 Results

As described above, we used a framework for text mining applications within R software named, tm: Text Mining Package [9], in data mining stage and, through it, we get the 20 most quoted words of each research line.

Figure 2 shows the R script used to mining data from the Knowledge Engineering applied to Electronic Government research line. The script uses a text mining library which produces the corpus formed form all the documents presented in the Knowledge Engineering applied to Electronic Government folder.

```
library(tm)
setwd('/BTD_EGC/ENGENHARIA DO CONHECIMENTO/Engenharia do Conhecimento Aplicada a Governo Eletrônico/')
docs <- Corpus(DirSource('./'))
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/|@|\\|")
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, removeWords, stopwords("portuguese"))
docs <- tm_map(docs, removeWords, c("conhecimento", "engenharia", "gestão", "mídia",
                                    "ser", "figura", "sobre", "forma", "the", "pode",
                                    "fonte", "quadro", "and", "cada","podem", "aae", "assim"))
docs <- tm_map(docs, stripWhitespace)
dtm <- DocumentTermMatrix(docs)
freq <- colSums(as.matrix(dtm))
ord <- order(freq)
freq[tail(ord, 20)]
```

Fig. 2. R Script, Corpus Analysis - Knowledge Engineering Applied To Electronic Government

After a series of transformations in the corpus (removing spaces, special characters, uppercases, numbers, punctuation, stop words and some words which are super frequent) is possible to applied a function which convert this corpus as a document term matrix. After this, is possible to count and show the most frequent words.

The taxonomic structure was constructed from these obtained words. However, an analysis process on this result was done, since many terms were constructed from two or more words.

The 322 documents found in the PPGEGC Theses and Dissertations Library Catalog was organized according to the area of knowledge and the research line to which they belonged, as described above. The Table 1 shows the result of the organization of the documents.

Table 1. Documents Organization of PPGEGC Theses and Dissertations Library Catalog

Concentration Area	Research Area	Number of Documents
Knowledge Engineering	Knowledge Engineering applied to Electronic Government	10
	Knowledge Engineering applied to Organizations	21
	Theory and Practice in Knowledge Engineering	46
Knowledge Management	Sustainability Knowledge Management	21
	Knowledge Management, Entrepreneurship and Technological Innovation	33
	Theory and Practice in Knowledge Management	95
Media and Knowledge	Media and Knowledge in Education	46
	Media and Dissemination of Knowledge	12
	Theory and Practice in Media and Knowledge	38

We find that the words resulting from the data mining process would not be enough to compose a concise taxonomy to describe all relevant terms to index the documents of the PPGEGC Theses and Dissertations Library Catalog.

The number of words found in the previous process was limited and did not represent all subjects researched in PPGEGC.

Therefore, we chose to include the keywords of the documents in the defined taxonomic structure. The keywords were obtained manually, completing the navigational taxonomic structure.

At the end of vocabulary survey stage we obtained a total of 583 terms that were organized by research lines and those ones, hence, were grouped by concentration area, as shown in Table 2.

Table 2. Quantity of raised terms

Concentration Area	Research Area	Number of Terms
Knowledge Engineering	Knowledge Engineering applied to Electronic Government	25
	Knowledge Engineering applied to Organizations	32
	Theory and Practice in Knowledge Engineering	74
Knowledge Management	Sustainability Knowledge Management	50
	Knowledge Management, Entrepreneurship and Technological Innovation	75
	Theory and Practice in Knowledge Management	142
Media and Knowledge	Media and Knowledge in Education	66
	Media and Dissemination of Knowledge	40
	Theory and Practice in Media and Knowledge	79

Figure 3 shows a subset form the taxonomy, representing the Knowledge Engineering applied to Electronic Government research line. This subset represents all the 25 words, emerged from the taxonomy building process.

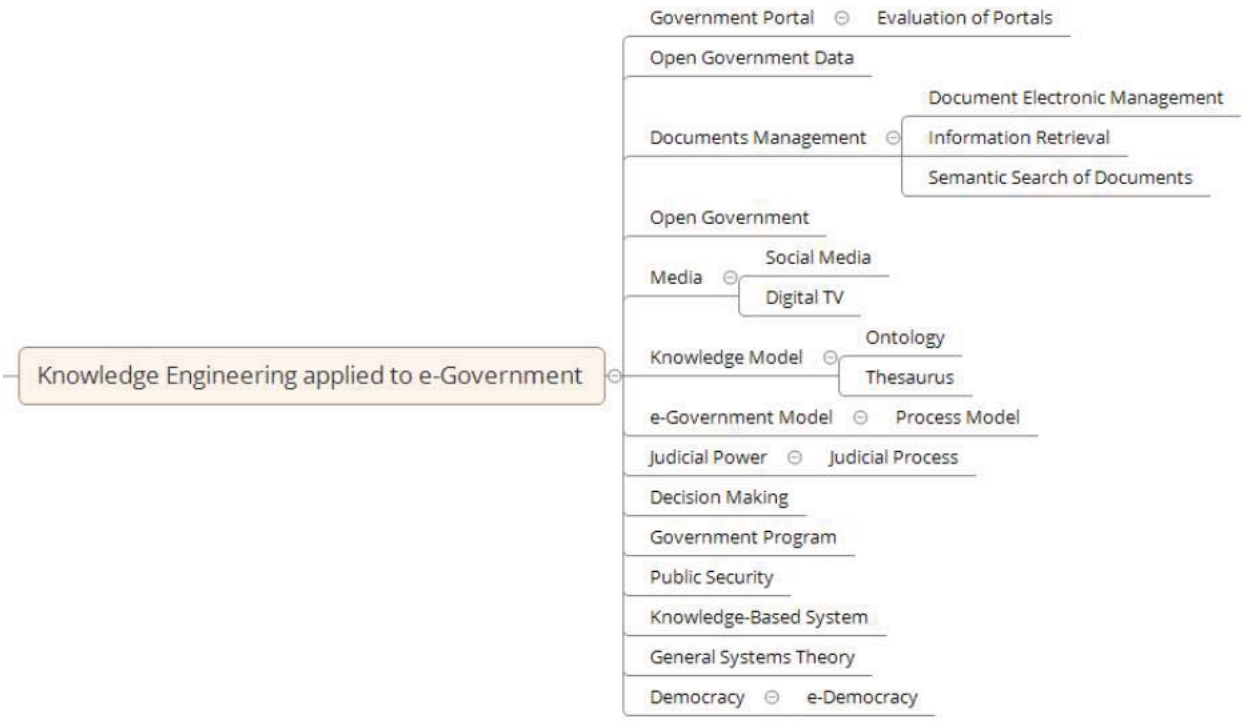


Fig. 3. Knowledge Engineering Applied To Electronic Government Hierarchical Navigational Taxonomy

6 Conclusions

Developing a navigational taxonomy presented in this paper, we proposed an indexing structure to organize the documents in the PPGEGC Theses and Dissertations Library Catalog, facilitating recovery of these documents. With this taxonomy the task of information retrieval from a specific domain is facilitated through a controlled vocabulary that expresses the main relevant terms of this particular area.

The proposed methodology for the extraction of information followed a semi-automated process which, at first, consisted of data acquisition (documents in .pdf format available at the PPGEGC Theses and Dissertations Library Catalog) and processing of such data, that is, the conversion of these documents into .txt format and encoding for standard UTF. This step was performed without many problems, and the codes and necessary controls were easily used.

The second stage, vocabulary survey, was carried out manually. In other words, it was necessary to analyze each of the 322 documents found on acquisition stage to generate the categorization (definition of which research line the thesis or dissertation better fit).

Many documents had indicated in its body text the research line, but some only had the concentration area and others offered no indication, being necessary to verify the associated tag in the bank of theses and dissertations, as well as all document content.

This stage took a long time and because of that, we suggest to PPGEGC to require all students to inform in the alignment with the program document subsection which area and research line the thesis or dissertation fits best.

The third stage, the mining of the most common terms, was performed automatically by R script that made changes in the corpus created for each research line, allowing the construction of the matrix of terms of each PPGEGC research line. This step required a curve of moderate learning by the authors, and the R language allows other analyzes in the corpus such as dendograms, word cloud, etc. that can be used in future work.

The matrix of terms frequently proved insufficient to develop navigational taxonomy, so the correlation of this matrix of each research line with the keywords of these documents on the line was essential. The survey of this vocabulary, set after this correlation, is the main contribution of this article.

Based on this taxonomy it is possible to think about the restructuring of PPGEGC Theses and Dissertations Library Catalog so that the search and recovery of these documents could bring a better user experience.

7 References

[1] Shenk, David. Data Smog: Surviving the Info Glut. Technology Review, v. 100, n. 4, p. 18-26, 1997.

[2] Hyman, H., Sincich, T., Will, R., Agrawal, M., Padmanabhan, B., & Fridy Iii, W. A process model for information retrieval context learning and knowledge discovery. Artificial Intelligence and Law, 23(2), 103-132, 2015.

[3] Jaime Lara, María de la Concepción Pérez de Celis and David Pinto. Dynamic Concept-Based Taxonomy used for image recovery based on their textual description . SemWeb09. 534(1): 28-36 (2009)

[4] PPGEGC. Histórico. 2015. Available: http://www.egc.ufsc.br/pos-graduacao/programa/historico/ [Acessed 19 March 2016].

[5] PPGEGC. Banco de Teses e Dissertações do EGC. 2016. Available: http://btd.egc.ufsc.br/ [Acessed 14 August 2015].

[6] Conway, Susan; SLIGAR, Char. Building taxonomies. Unlocking knowledge assets, n. s 1, p. 105-124, 2002.

[7] Garshol, Lars Marius. Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. Journal of information science, v. 30, n. 4, p. 378-391, 2004.

[8] Francis, Louise; FLYNN, Matt. Text mining handbook. In: Casualty Actuarial Society E-Forum, Spring 2010. 2010.

[9] Meyer, David; HORNIK, Kurt; FEINERER, Ingo. Text mining infrastructure in R. Journal of statistical software, v. 25, n. 5, p. 1-54, 2008.

An analysis of the implications of Maslow's Hierarchy of Needs for networked learning design and delivery

Jonathan Bishop
Centre for Research into Online Communities and E-Learning Systems, Swansea, Wales, GB

Abstract - *The Hierarchy of Needs proposed by Abraham Maslow has been adopted by many groups of practitioners as a way to understand their customers and users. It argues that there are universal human needs, namely physiological, security/safety, social, self-esteem/ego and self-actualisation. Maslow contests that unless the former of these are met, the latter cannot be. This paper demonstrates the need for the continual review and modification of teaching and learning plans to meet the changing needs of learners, which in this case relates to considering the impact of networked learning.*

Keywords: Abraham Maslow, hierarchy of needs, education theory, networked learning, learning stratefies

1 Introduction

These The Hierarchy of Needs theory provided by Abraham Maslow [1, 2] has been argued to be one of the simplest ways of understanding human behaviour [3], but in the digital age a greater degree of complexity is needed to understand the place humans in contemporary societies [4]. Maslow's model is a popular motivation theory in social science, but has been criticised as untestable with insufficient empirical evidence [5]. As can be seen from Fig. 1, Maslow's hierarchy is often illustrated in the shape of a pyramid, with the principle that the 'deficit needs' that are most essential to life are at the bottom, and the 'esteem needs,' which he deems least essential are at the top [6].

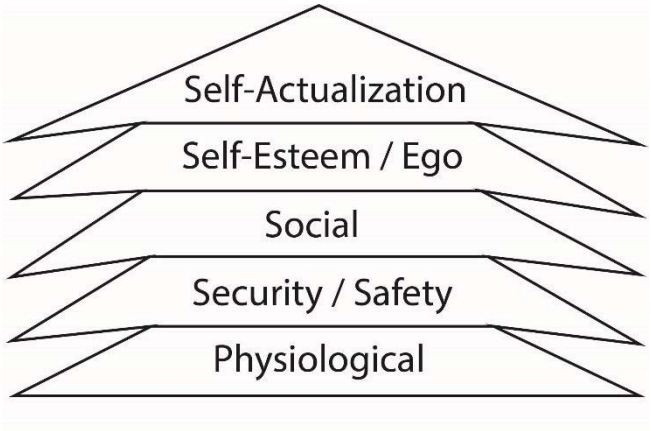


Fig. 1 Maslow's Hierarchy of Needs

It has been said that in online environments that Maslow's hierarchy cannot apply, especially in the case of those with digital addiction, as they will often give up their so-called basic needs to focus on their so-called esteem needs [7]. Equally it has been found that in intense offline environments where people are absent of basic needs, such as concentration camps, the sense of solidarity and compassion between humans associated with the esteem needs is still evident [8].

Even so, with current government policy around the provision of free school meals to improve learning outcomes [9-12] and the drive for greater use of ICTs in education at all levels [13], including where distance learning might be preferred [14], it is clear Maslow's hierarchy deserves another inspection.

Networked learning as a concept that has existed since the 1990s, which refers to new ways of using information systems to influence teaching and learning [15]. A means of enabling e-learning, networked learning makes use of not only knowledge around computer science, but social science also, which jointly put in in the context of information systems and human-computer interaction. The contribution of motivational theories to learning design in such environment has been significant [16, 17]. Indeed, Scholars in this area have said the 'hierarchy of needs,' proposed by Abraham Maslow [1, 2], is suited to the design of information systems because it is "orderly" [18]. Others have said it is suitable for the design of online communities [19, 20], and others still have said it has application in gaming in terms of understanding "clans" [21]. The societal implications of such 'received wisdom' needs to be verified in order to prevent the development of erroneous practice due to misinterpretations in Maslow's hierarchy, including those that have resulted from scholars mixing and matching the original model for their own purposes. It has already been shown prior to the advent of social media that this hierarchy is unsuited to virtual environments [7]. Even though this is almost without doubt, this paper takes another look at Maslow's hierarchy of needs to assess its suitability for design and delivery of learning in networked learning environments, and educational contexts in general.

2 The development of learning theory from psychological traditions

Please

This section explores the complexities of the psychological foundations of teaching and learning on the basis that Maslow's

hierarchy of needs cannot be considered in isolation of the historical and contemporary psychological developments in the context of practice that existed the time it was formed, nor those contemporary practices that exist today. By critically analysing the relationship between competing psychological perspectives in the context of learning with Maslow's hierarchy, then is role in contemporary learning environments can be defined.

It has been argued that when one thinks about educational psychology in an historical context, one should focus on the scientific and empirical study of education [22]. According to Abraham Maslow himself, new developments in psychology have always generated a revolutionising many theoretical developments in psychology have been later found to have been based on insecure empirical foundations [23], something that will be discussed in this article in relation to Maslow's hierarchy of needs.

Maslow's work is not the only one that has considered the impact of the environment on intrinsic motivations. Gestalt psychology, for instance, is based on the premise that the mind translates external stimuli holistically rather than the sum of their parts [24]. On the other hand, behaviourist psychology suggests the opposite [25, 26]. The battle between Gestalt and behaviourist visions of psychological function and competence marked the early years of 20th century, could be considered to be continuing to do so, albeit in modified ways [27]. Whilst behaviourism has had some traction in the design of networked learning systems [28], it is now known that human understanding of reward systems is much more complex [29].

In the past, it had been argued that the learning processes of insight and intellectual development of humans is best explained with Gestalt theory, with behaviourism being more suited to understanding conditioning, such as in contemporary education and training [30]. However, following on from the gestalt-behaviourist battle was the advent of cognitivism, which attempted to resolve the empirical deficit in these theories [31]. Behaviourism depended on a process of trial and error by the person learning was aimed at, with rewards and punishment following as part of the process, lacking objectivity and systematic reliability [30]. However, many of these ignore the role of human memory studies in the learning process, which find stimuli-response connections are more cognitive than was thought by behaviourist [32].

2.1 Maslow's Hierarchy of needs in comparison with other models

Even though Maslow's hierarchy of needs has been regarded to be an important part of introducing motivational theories into education, with it still being popular today [5, 33], but it has consistently been unsupported by evidence from the research community, with many saying it cannot be put in practice or verified by empirical research [5, 33]. Indeed, even Maslow himself a decade after the first publication of his theory, expressed reservations about the application of it to work motivation and similar contexts [34].

3 Limitations of the Hierarchy of Needs for design and delivery of networked learning

It has been argued that Maslow's Hierarchy of Needs theory can form part of teaching, learning social strategies in diverse occupational settings, including among online community managers [19, 20, 35, 36]. As can be seen from Table 1, Kim has linked essential user requirements in using online communities to the five levels of the Hierarchy of Needs.

Table 1 Kim's linking between Maslow's hierarchy of needs and user requirements

Hierarchical Stage	Manifestations in online environments
Physiological	System access: the ability to maintain one's identity, and participate in a Web community
Security and Safety	Protection from hacking and personal attacks, the sense of having a "level playing field"
Social	Belonging to a community as a whole, and sub-groups within the community
Self-esteem	The ability to contribute to the community, and be recognised for those contributions
Self-actualization	The ability to take on a new community role that develops skills and opens up new opportunities

The advent of social media as a form of mass communication has the opinions of those once in the public sphere controlled by the media, now form part of the 'public square' where they can self-publish [37, 38]. With networked learning concepts like Classroom 2.0 [12, 39-41], the suggestions of Kim in Table 1 seems dated when it comes to devising strategies to promote teaching and learning in online environments.

A weakness of Maslow's hierarchy of needs is that it assumes everyone has the same drivers, meaning it therefore lacks the capacity to take account of individual differences. As has been shown in research on the Classroom 2.0 initiative [12, 39-41] educators and e-learning systems need to demonstrate an awareness of the preferences of learning in the learning situation. Such e-learning systems can take account of the various social contexts in learning environments, including varying intelligences. For instance, it has been shown that e-learning systems can be adapted to take account of attainment levels so that the content delivered is at an appropriate level of challenge [40]. Such approaches can be seen as justified because they allow for each learner to have their individual differences accommodated in a way not provided for in Maslow's model. Learning activities used through Classroom 2.0 can include interacting with learners from other educational establishments, taking part in a class where generic questions are tailored to each person's interests and abilities [40]. If Maslow's model were to be applied here, then it would suggest that a learner would be unable to be motivated away from the

more solitary tasks. However, the intention of the Classroom 2.0 approach to network learning is that the self-esteem generated in the abler learners will result in them helping the less able learners after they have completed their own work. A problem of mixed ability classes has been that the abler speed ahead in group tasks, leaving the less able falling behind. This problem does not exist with Classroom 2.0, as each student has their own terminal, where they can rely on the system to take account of their differences, meaning the abler are occupied and do not take over.

4 Changes to learning and instruction needed in the digital age

The Classroom 2.0 approach to networked learning involves the joining up of each learner's device and provides a custom experience, is not the only contemporary mode of delivering teaching and training that challenge's Maslow's hierarchy. New approaches to learning, using tablet devices like iPads, mean the relevance of Maslow's hierarchy and other traditional theories of motivation is diminishing.

The distinction between what Maslow calls 'physiological' and 'social' is becoming less relevant. Those part of Generation Next are going through school at a time where they are multitasking not by sitting in front of a PC, but by walking and using their smart device at the same time, meaning their physical coordination will be as strong as their capacity to communicate, unlike with earlier generations [42]. The behaviourist and cognitivist approaches to the design of learning and construction of learning plans are therefore irrelevant, and post-cognitivist ones based on accounting for environmental and ecological factors are therefore needed [43]. Maslow on the other hand argues that the environment influences only one's physiology and sense of security. It is in fact the case that the extent to which one feels capable of being social and having a sense of self-esteem is dependent on one's surroundings.

However, it has been argued that even if behaviourist reinforcement theories have been discredited, this does not mean that stimulus-response learning understood through these theories are non-existent [44]. Indeed, even in post-cognitivist models, which represent latest thinking, a stimulus is what a human receives and a response what they produce [17].

4.1 Changes needed to teaching and learning strategies and modes of delivery

It has been established by this paper that Maslow's hierarchy of needs is not suited as a teaching and learning strategy in a range of online contexts, such as networked learning. This section will explore how the benefits sought by those who like Maslow's model for its simplicity [18], can be achieved with a more recent post-cognitivist model design for the digital age, namely ecological cognition [17, 45]. To do this the use of a range of 'e-tivites' [46] will be deployed and evaluated, including their advantages and disadvantages as delivery modes in contrast with those typically used to apply Maslow's hierarchy. Whilst Gilly Salmon says there are 5 levels of providing e-tivities this section will focus on the first, namely 'access and motivation.' The reason for this is that the original basis of Maslow's hierarchy was to understand human motivation. The processes used in Transactional Analysis, namely activities, and the "Learn, Create, Communicate" (LCC) approach to teaching [40, 47, 48], will also be used to show how these occur in specific settings.

4.1.1 Activities for the "Learn" stage

The learn component of the LLC model advocates that acquiring knowledge should be the first stage of taking on a new topic [40, 47, 48]. Relevant e-tivities at this stage are ice-breakers, quizzes, and the 'my brand' activity [46].

"Ice-breakers" are activities for encouraging participation of newcomers to a course, which involves increasing their involvement over two to three weeks [46]. A process known as 'delurking,' these activities encourage users who might be afraid of saying something wrong or not fitting in – called lurkers – to get the confidence to take part [49-51]. Lurking in the context of Transactional Analysis can be seen to be the activity of 'Withdrawal' [52] and delurking is the process of avoiding or reversing such withdrawal. Whilst organisations deploying Maslow's hierarchy use ice-breakers in their training, the model has been criticised for its usefulness in this regard because of how it focuses on the individual existing outside of a group [53]. Ice-breakers are inherently social activities [54], meaning Maslow's model is unsuitable as it assumes being social is not an essential part of learning, but additional.

The "Quiz" form of e-tivity is where learners are asked to tell others about themselves, such as their job and personal interests, and in some cases a prize is awarded for contributions [46]. This is evident in the TA activity of 'rituals,' which is where participants know the socially acceptable things to say and what not to [52]. Quizzes are known to help in learning environments based on Maslow's hierarchy, where they have been used throughout courses to engage difficult to educate groups [55]. Despite the fact that quizzes are usually done on an individual basis, one might question whether the lack of priority of being social in Maslow's model would make learners more goal-driven than they need to be when doing informal quizzes.

The "My brand" e-tivity invites learners to say why they choose certain brands and what it says about them [46]. This reflects the TA activity of 'playing,' because it is about generating ideas as opposed to selecting and excluding robustly [52].

4.1.2 Create

In terms of the 'Create' stage of the LCC model, it is assumed that acquiring knowledge on its own does not mean one shows mastery of the topic, as that knowledge needs to be put into action [40, 47, 48]. To do this online, the e-tivities of 'Images' and Talents' are relevant [46].

With regards to the e-tivities for "Images", it is recommended that learners be invited to share pictures of themselves and their lives by linking to them on the Internet [46]. This is reflected in the TA activity, 'closeness' [52]. The reason for this is that at this stage, learners will have developed sufficient trust to be able to share true feelings and express views candidly [52].

Such 'show and tell' activities are said to promote what Maslow calls 'self-actualisation' as they are meant to help learners develop a psychology of being [56]. It is clear that such activities go beyond one's personal development, showing an appreciation of others also.

An e-tivity called, "Talents," involves learner having an imaginary sum of money and using e-commerce websites to choose the products they would most want, which could be related to the course [46]. This can be seen to be reflected in the TA activity of 'working,' where is where learners participate in order to achieve things together, including reviewing ideas. Maslow's model is not suited to situations where togetherness is a factor [57], with some arguing that it is important at all stages of development, especially in relation to physiological need around, whereas Maslow puts it as an esteem need and therefore not essential [58].

4.2 Communicate

The communicate stage of the LLC model suggests that learning is a social process and that without expressing one's learning it can be difficult to retain it [40, 47, 48]. A number of e-tivities can help with this stage, namely wanderlust and hall of mirrors.

The "Wanderlust" e-tivity involves the educator posting a link to a location relevant to the course and asking learners to post links to other locations that either reflect where they are from or which they think relates to the course in some way [46]. This is reflected in the TA activity of 'pastimes,' as it involves learners discussing things in an informal and more social way, with little reference to the serious elements of the course [52]. In such a context, especially where the course has learners from many cultural backgrounds, it has been argued that Maslow's hierarchy is unsuitable as it only fits one culture of mainly successful White men [59]. Whilst talking about physiological needs might be suitable to those from impoverished backgrounds, to expect others to be forced to fit Maslow's model, would do little to help assist teaching and learning.

The e-tivity named, "Hall of mirrors," involves the educator posting links to five websites reflecting something about the course being taught and invites users to comment on their similarity, why they might use them, which they would be most likely to use, and their emotions using them [46]. This too can be seen to reflect the TA activity of pastime, as learners literally 'pass the time' [52], discussing what they learned on the course in reference to other sources of information. It is argued that by having self-actualisation, humans have a sense of spirituality so that a sense of community is enhanced [60]. However, it is known that learners often have time-pressures, meaning expecting post-class interactions might be unlikely [61].

5 Discussion

This paper has demonstrated comprehension of the need for the continual review and modification of teaching and learning plans to meet the changing needs of learners. This has been done by showing how Maslow's hierarchy of needs is not appropriate in the digital age, where networked learning environments, such as those based on Classroom 2.0, are a core part of teaching and learning.

To assess Maslow's model effectively, its impact on 'e-tivities' (i.e. activities that take part in electronic environments) was discussed, along with how they relate to those activities in Transactional Analysis (TA). This investigation found that Maslow's hierarchy is not suitable as a means to support teaching and learning in networked learning environments, because it puts too much focus on the individual. It was found that time was more of a pressure than deficit needs such as food and safety, and to expect learners to focus on discussing these latter factors during their online learning is not culturally appropriate.

The paper concludes that Maslow's hierarchy 's claim that the drive to be social is a higher one over physiological is not accurate. In networked learning environments social interaction is by far the most important factors that motivates participation, as the evaluation into e-tivities showed.

6 References

[1] A. H. Maslow. "A theory of motivation"; *Psychological review,* 50., 4, 370-396, 1943.

[2] A. H. Maslow. "A theory of human motivation"; *Readings in managerial psychology,* 20, 1989.

[3] Azilah Kasim, Hisham Dzakiria, Chansoo Park, Nor Azila Mohd Nor, Mohamad Fawzi Mokhtar & Rashid Radha, Jasmine Raziah Radzi Radha. "Predictors of travel motivations: the case of domestic tourists to island destinations in northwest of Malaysia"; *Anatolia,* 24., 2, 188-205, 2013.

[4] Donald Arthur Norman. "Living with Complexity". MIT Press, 2010.

[5] Kee Mun Wong & Ghazali Musa. "Retirement motivation among 'Malaysia My Second Home'participants"; *Tourism Management,* 40., 141-154, 2014.

[6] Nik Ahmad Sufian Burhan, Mohd Rosli Mohamad, Yohan Kurniawan & Abdul Halim Sidek. "National intelligence, basic human needs, and their effect on economic growth"; *Intelligence,* 44., 103-111, 2014.

[7] Jonathan Bishop. "Increasing participation in online communities: A framework for human–computer interaction"; *Computers in Human Behavior,* 23., 4, 1881-1893, 2007.

[8] Mahmoud A. Wahba & Lawrence G. Bridwell. "Maslow reconsidered: A review of research on the need hierarchy theory"; *Organizational Behavior and Human Performance,* 1976., 15, 212-240, 1976.

[9] L. Moore, G. F. Moore, K. Tapper, R. Lynch, C. Desousa, J. Hale, C. Roberts & S. Murphy. "Free breakfasts in schools: design and conduct of a cluster randomised controlled trial of the Primary School Free Breakfast Initiative in Wales [ISRCTN18336527"; *BMC public health,* 7., 258, 2007.

[10] Katy Tapper, Simon Murphy, Laurence Moore, Rebecca Lynch & Rachel Clark. "Evaluating the free school breakfast initiative in Wales: methodological issues"; *British Food Journal,* 109., 3, 206-215, 2007.

[11] M. Nelson, K. Lowes & V. Hwang. "The contribution of school meals to food consumption and nutrient intakes of

young people aged 4–18 years in England"; *Public health nutrition,* 10., 07, 652-662, 2007.
[12] Jonathan Bishop. "Microeconomics of Education and the Effect of Government Intervention: The Role of Classroom 2.0 in Facilitating the UK Government's Schools Policies"; *Transforming Politics and Policy in the Digital Age* (IGI Global) Jonathan Bishop (Ed.), 39-512014.
[13] Neil Christopher Charles Brown, Michael Kölling, Tom Crick, Simon Peyton Jones, Simon Humphreys & Sue Sentance. "Bringing computer science back into schools: Lessons from the UK". Proceeding of the 44th ACM technical symposium on Computer science education, ACM, 2013. , 269-274.
[14] Elizabeth Locke & Sally Hewlett. "Large Scale Power Generation: Up-Skilling Welsh Industry"; *Energy and Environment Research,* 4., 1, p74, 2014.
[15] Irene Hanraets, Joitske Hulsebosch & Maarten de Laat. "Experiences of pioneers facilitating teacher networks for professional development"; *Educational Media International,* 48., 2, 85-99, 2011.
[16] J. Stephenson. "Teaching and Learning Online: New Pedagogies for New Technologies". Routledge Falmer, 2001.
[17] Jonathan Bishop. "Ecological Cognition: A New Dynamic for Human-Computer Interaction"; *The Mind, the Body and the World: Psychology after Cognitivism* (Imprint Academic) Brendan Wallace, Alastair Ross, John Davies & Tony Anderson (Eds.), 327-3452007.
[18] B. Shneiderman. "Leonardo's Laptop: Human Needs and the New Computing Technologies". MIT Press, 2002.
[19] Amy Jo Kim. "Community Building on the Web: Secret Strategies for Successful Online Communities". Peachpit Press, 2000.
[20] Jonathan Bishop. "Development and Evaluation of a Virtual Community". 2002. .
[21] Michael Del Grosso. "Design and Implementation of Online Communities.". 2001. .
[22] David C. Berliner. "Educational psychology: Searching for essence throughout a century of influence"; *Handbook of educational psychology,* 2., 3-42, 2006.
[23] Andrew W. Ellis. "Modality-specific repetition priming of auditory word recognition"; *Current Psychological Research,* 2., 1-3, 123-127, 1982.
[24] Shaun Foster & David Halbstein. "3D Design and Photogrammetry"; *Integrating 3D Modeling, Photogrammetry and Design* (Springer) Anonymous 69-962014.
[25] Ivan P. Pavlov. "Conditioned reflexes". Routledge and Kegan Paul, 1927.
[26] B. F. Skinner. "The Behavior of Organisms: An Experimental Analysis". Appleton-Century-Crofts, 1938.
[27] Gavin Kendall & Mike Michael. "Thinking the unthought:: towards a Moebius strip psychology"; *New Ideas in Psychology,* 16., 3, 141-157, 1998.
[28] Geneen Stubbs & Mike Watkins. "Re-engineering CBL development". Proceedings of 26th Annual Conference on Frontiers in Education 1996 (FIE'96), IEEE, New York, NY, 1996. , 1387-1390.
[29] BISHOP, J., Ed. 2014. Gamification for Human Factors Integration: Social, Educational, and Psychological Issues. IGI Global, Hershey, PA.
[30] Leopold Bellak. "A note on some basic concepts of psychotherapy"; *The Journal of nervous and mental disease,* 108., 2, 137-141, 1948.
[31] Ed Burton. "Representing representation: Artificial Intelligence and drawing"; *Computers & Art,* 2002.
[32] L. Cahill, J. L. McGaugh & N. M. Weinberger. "The neurobiology of learning and memory: some reminders to remember"; *Trends in neurosciences,* 24., 10, 578-581, 2001.
[33] Jay J. Hochstetler. "Revising the Volunteer Functions Inventory: An Exploratory Study of Additional Functions". 2013. .
[34] Gary P. Latham. "Work motivation: History, theory, research, and practice". Sage publications, 2011.
[35] Amy Jo Kim. "Community Building on the Web: Secret Strategies for Successful Online Communities". 2006.
[36] Jonathan Bishop AND Lisa Mannay. "Development and Evaluation of an Adaptive Multimedia System". 2002. .
[37] D. Tapscott & A. D. Williams. "Macrowikinomics: Rebooting Business and the World". Atlantic Books, 2010.
[38] D. Tapscott & A. D. Williams. "Wikinomics: how mass collaboration changes everything". Atlantic Books, 2006.
[39] Gabriella Taddeo & Simona Tirocchi. "Learning in a "Classi 2.0" Classroom: First Results from an Empirical Research in the Italian Context"; *Didactic Strategies and Technologies for Education: Incorporating Advancements* (IGI Global) Paolo M. Pumilia-Gnarini, Elena Favaron, Elena Pacetti, Jonathan Bishop & Luigi Guerra (Eds.), 57-672012.
[40] Jonathan Bishop. "Cooperative e-learning in the multilingual and multicultural school: The role of 'Classroom 2.0' for increasing participation in education"; *Didactic Strategies and Technologies for Education: Incorporating Advancements* (IGI Global) P. M. Pumilia-Gnarini, E. Favaron, E. Pacetti, J. Bishop & L. Guerra (Eds.), 137-1502012.
[41] Jonathan Bishop. "The Persuasive and Assistive Interaction Extension (PAIX): A position paper on using gamified behavior management systems for reducing flame trolling in schools based on Classroom 2.0". The 13th International Conference on Internet Computing (ICOMP'12), Las Vegas, NV. 16-19 July 2012, WORLDCOMP, Las Vegas, NV, 2012. .
[42] Sherry Turkle. "Alone Together". Basic Books, 2013.
[43] B. Wallace, A. Ross, T. Anderson & J. Davies. "The Mind, the Body and the World: Psychology After Cognitivism?". Imprint Academic, 2007.
[44] Bertram Gawronski & Galen V. Bodenhausen. "Theory evaluation"; *Theory and Explanation in Social Psychology,* 1-12, 2014.
[45] Jonathan Bishop. "The Role of Multi-Agent Social Networking Systems in Ubiquitous Education: Enhancing Peer-Supported Reflective Learning"; *Multiplatform E-Learning Systems and Technologies: Mobile Devices for Ubiquitous ICT-Based Education* (IGI Global) T. T. Goh (Ed.), 72-882009.
[46] Gilly Salmon. "E-tivities: The Key to Active Online Learning". RoutledgeFalmer, 2003.

[47] R. Agostini. "Technology of Education and Music Teaching: New Responses to Old Issues"; *Didactic Strategies and Technologies for Education: Incorporating Advancements* (IGI Global) P. M. Pumilia-Gnarini, E. Favaron, E. Pacetti, J. Bishop & L. Guerra (Eds.), 27-372012.
[48] Jonathan Bishop. "Lessons from The Emotivate Project for Increasing Take-up of Big Society and Responsible Capitalism Initiatives"; *Didactic Strategies and Technologies for Education: Incorporating Advancements* (IGI Global) P. M. Pumilia-Gnarini, E. Favaron, E. Pacetti, J. Bishop & L. Guerra (Eds.), 208-2172012.
[49] PFAFFENBERGER, B., Ed. 1995. Que's Computer & Internet Dictionary. Que Corporation, Indianapolis, IN.
[50] John Cowpertwait & Simon Flynn. "The Internet From A to Z". Icon Books Ltd, 2002.
[51] Jennifer Preece, Blair Nonnecke & Dorine Andrews. "The top 5 reasons for lurking: Improving community experiences for everyone"; *Computers in Human Behavior,* 2., 1, 42, 2004.
[52] Julie Hay. "Transactional Analysis for Trainers: You're Guide to Potent & Competent Applications of TA in Organisations". Sherwood Publishing, 1996.
[53] Richard Heslop. "'Doing a Maslow': Humanistic Education and Diversity in Police Training"; *The Police Journal,* 79., 4, 331-342, 2006.
[54] Rodney P. Beary. "Inquiring trainers want to know"; *Training and Development,* 29., 22-25, 1994.
[55] Rebecca S. Watts. "An Alternative School within a School: A Case Study on Meeting Motivational, Curricula, and Instructional Needs of At-Risk Students.". Anual Meting of the Mid-South Educational Research Asociation, Bowling Gren, KY. 15-17 November 2000, ERIC, 2000. , 1-27.
[56] Kathryn Katzman Rolland. "Show and Tell: Developing an Appreciation of Diversity"; *Journal of Health Education,* 24., 2, 116-117, 1993.
[57] Vladimiras Grazulis. "Personnel Management and Vitality Phenomenon of A. Maslow‘s Theory of Needs"; *Human Resources Management and Ergonomics,* 2, 14-21, 2007.
[58] Charles D. McDermid. "How Money motivates Men"; *Business horizons,* 3., 4, 93-100, 1961.
[59] Casey P. Hayden. "A hierarchy of needs in international relations". 2009. .
[60] Christopher P. Neck & John F. Milliman. "Thought self-leadership: Finding spiritual fulfilment in organizational life"; *Journal of Managerial Psychology,* 9., 6, 9-16, 1994.
[61] Norah Jones, Haydn Blackey, Karen Fitzgibbon & Esyin Chew. "Get out of MySpace!"; *Computers & Education,* 54., 3, 776-782, 2010.

Role of Information Communication Technology in Modern Educational System at University Level

Nouman Maqbool Rao[1], **Shazia Kanwal**[2,] **Muhammad Yasir Ali Abass**[2,]
[1]P&D, Punjab Higher Education Commission, Lahore, Pakistan
[2]Salford Business School, University of Salford Manchester Lancashire United Kingdom
[2]Management Sciences, Lasbela University of Agri, Water & Marine Sciences, Uthal, Pakistan

Abstract: *Technological storm has significant impact on our lives. People, as whole have become so dependent on technology that they would not know how to survive without it. The digital revolution is driving the societal trends. Likewise, technological advancements have tremendous influence on educational institutions as well. The substantial growth in the demand of higher education has generated the need of innovative and flexible approaches of learning. In an educational system "teaching" and "learning" are the two major activities besides "assessment" which is a coordinating activity. The Information Communication Technology (ICT) has a potential to transform the different areas of the educational system. In this paper, our focus is to identify the current issues and challenges in our educational system and to propose the role of Information Communication Technology and its successful implementation to overcome these challenges. This paper also provides some recommendations which could be used as a catalyst for the promotion of information communication technology services in the context of higher education.*

Keywords*:* ICT, Teaching, Learning, Simulation, Education, Assessment

1. Introduction

Although information technology is not a panacea for all of the shortfalls associated with our educational system; it offers the potential not only for significantly enhancing learning for all learners but also for transforming the way we learn [4]. An educational system is characterized by two fundamental activities i.e. teaching and learning. These activities involve huge resources of both state as well as public with the ultimate objective of effective deliverance. The role of the assessment system in this perspective is to co-ordinate the teaching and learning activities and to evaluate the system performance.

The main objectives of an educational system; however its functional domain also includes services which it has to offer to its constituent elements [3]. Therefore, in general educational system have the following activities:

a) Teaching
b) Learning
c) Assessment and
d) Other Services

The conventional education system, particularly its assessment system is awfully strained and has virtually limited its functioning to the conduct of examinations and declaration of results. This has significantly affected the performance of the entire system. Presently, the studies argued that the role of ICT in the educational sector is limited to the deliverance of services and few more activities, whereas the challenges are multifarious. Although the advancement in information and communication technologies have the potential to enhance lifelong learning [2].

These technologies can be used to address the changing demands of the sector:

- for more flexible learning;
- for expansion of educational services to national and international markets; and
- for more cost-effective delivery of education and related services in an increasingly competitive environment.

2. System Challenges

The four different areas of activities identified above are in a state of extreme dynamics and the conventional system is unable to cope up the challenges associated with new opportunities and developments. Some of the main challenges in each of the activities of an educational system are as follows:

2.1 Teaching:

ISBN: 1-60132-441-3, CSREA Press ©

The Challenges with the teaching include:

- ✓ To plan the lecture in a highly efficient manner such that both teacher and learner participation become mandatory.
- ✓ To provide a suitable environment for efficient transformation and dissemination of vital information
- ✓ To encourage the student community with the concept of "learn by doing."
- ✓ Making the most of the talents of the students, irrespective of their physical and mental disabilities.

2.2 Learning:

The learning scenario is also facing extreme dynamics and some of the main challenges associated with it are as under:

- ✓ Providing information and knowledge anywhere, anytime, anyway and anyhow.
- ✓ Allow more flexible access to education by reducing barriers of time and place of study.
- ✓ Acquire new skills in a way that is compelling and engaging.
- ✓ Participate in networked and face-to-face communities of learners, composed of teachers, mentors, domain experts, and "cognitive" tutors that collectively approach the effectiveness of a one-on-one human tutor.
- ✓ Using simulation for problem-solving approaches.
- ✓ Barrier of languages is made irrelevant

2.3 Assessment:

The deliverance of any education system is largely dependent on its examination system (assessment). The development of an assessment system which can meet its prime objectives of achieving desired validity and reliability is a real challenge at the global level. Some of the challenges in an assessment system are:

- ✓ Continuous Assessment Process
- ✓ Design the Assessment System in such a way to achieve the desired Validity, Reliability, and Transparency.
- ✓ Receive continuous and meaningful feedback of assessment.
- ✓ Make it comprehensive enough to explore the potential of the candidates.
- ✓ Make it healthy enough to develop higher order skills of comprehension in students.
- ✓ Make it student friendly exercise for every topic covered.
- ✓ Develop it as an integral part of the education system to co-ordinate the teaching & learning process.
- ✓ Design it in such a way which forces comprehensive reading.

2.4 Other Services:

The deliverance of service has been an essential element of an educational system. In order to make the system efficient in all spheres, the services have to be provided to all its users. Following are some of the main challenges in services:

- ✓ Tap into rich, universally accessible digital libraries with books, articles, material, and data sets.
- ✓ Dynamic administration of record and exchange information
- ✓ Develop information systems which support all the users of the system for their necessities.
- ✓ Provide services which ensure delivery of information anytime and anywhere at affordable costs.

3.0 Scope for ICT

The progression in the computer programming has reached a level wherein backed up by the massive databases artificial intelligence is incorporated into the systems. The Research Challenge is to provide learning environments that approach the characteristics as listed above [3]. Such systems, properly used, can produce a significantly better-educated populace by combining advances in learning sciences for human resource development with advances in information technology. The task of developing the desired system involves a number of technical challenges in the following distinct, but interrelated areas:

3.1 Cognitive Tutors

The development of a perfect human tutor involves significant resources in terms of both times as well as wealth. The development of a machine based tutor with the present level of computer programming has become a reality. It has been observed that it is possible for an automated tutor to improve student performance by roughly one standard deviation from the mean for some high school mathematics students. This is a dramatic result. One reason that such tutors are not widely available is because the significant human effort is required to develop the specialized knowledge base for each different subject. In addition, we do not

ISBN: 1-60132-441-3, CSREA Press ©

understand fully the conditions under which such tutor will be effective. A significant progress must be made in crafting knowledge representations that are both interoperable and reusable. We need to develop models of the various styles in which a student learns, as well as appropriate pedagogies and assessment techniques.

The knowledge representations that underline such tutors should also be designed to incorporate new knowledge about a subject area, as well as advances in knowledge and techniques associated with both pedagogy and assessment.

3.2 Simulation-based teaching

Multimedia has acquired great significance in teaching and learning. It has a tremendous scope for both teaching as well learning even in the most complex systems. An important genre of next-generation educational software, particularly for training scientists, mathematicians, engineers, and technologists, is what might be called a clip model. By loose analogy to the well-known galleries of copy-and-paste 2-D clip art, a clip model is an interactive microworld, typically simulation- or rule-based, that expresses both geometry and behavior of the modeled entity or concept. It is a self-contained object ready to be embedded in a context such as a hypermedia learning module. Firstly, they are designed to be combined to form larger models, for example, a heart model may be connected to a vascular system model and to a lung model to create a composite model that simulates respiration and oxygenation of the blood as it is distributed throughout the body. Second, no single model suffices for all learning purposes. Perhaps dozens, if not hundreds, of heart models are needed to meet the needs of learners at different levels of understanding and with different kinds of backgrounds and learning styles.

3.3 General Purpose Online Assessment System

It has been observed that almost all the areas of the assessment system require major reforms to make it as an effective part of the educational system. The Information Communication Technology has a tremendous scope for use in many operations in an Examination System besides the compilation of results and other student based services. In fact, most of the organizations have already switched over to a computerized system of a compilation of results and even Management Information system supporting the Management in running the system efficiently are already available. Similarly, students related information system providing relevant information to the students also exists in Organizations but these support systems do not strengthen the basic design issues and the objectives of the assessment system. In our opinion, support solutions are required to be developed for all key entities (teachers, students, examination body) involved in the system. Similarly, activities like question paper setting, evaluation, the conduct of exam etc require being supported by the technology to ensure the development of quality assessment system.

One of the essential components of the teaching-learning process is encouraging and creation of the system of self-studies among the students. In this context, teachers have been traditionally giving assignments to the students and it serves the purpose only once each student is given a different assignment and assessed once it is submitted. With the increase in the number of students, assigning different assignments to students and then evaluating each of them is an issue. This unit of the system requires being modernized by the support of the technology. In this regard, we propose that the development of an E-assignment system which shall assign a different assignment to each of the students and assess them automatically for their assignment on some other scheduled day without adding any additional burden to the teacher.

3.4 Learning in context/just-in-time learning

It is crucial that each individual learner is able to learn in the physical, social, cultural, and virtual context which is appropriate to that learner. In addition, the learner may have a more-focused objective for a particular learning session. Just-in-time learning over such a broad range of contexts presents a number of technical challenges, including the support of reliable and ubiquitous computing, access and control of remote instruments, flexible digital object sharing, and user interfaces for small-format mobile devices. The automated teacher must tailor delivered information to fit that context and the goals of the just-in- time learner.

4.0 Recommendations of ICT in Higher Education

The various recommendations for which have been proposed for using ICT practices in the field of education are given below

- ✓ Reliable Internet Connectivity is needed to facilitate the students with ICT benefit and ensure its upkeep.
- ✓ Accessibility to the ICT facilities by providing hardware/ software in the form of laptops, desktop, PDA's and other allied devices.

ISBN: 1-60132-441-3, CSREA Press ©

- ✓ Virtual/ Smart Classrooms need to be established in all Higher Educational Institutions.
- ✓ Availably of Competent Human Capital with relevant skill sets.
- ✓ ICT must be integrated with curriculum to make it competitive and bring it in tune with contemporary requirements.
- ✓ Allocation of funds by the Government for developing infrastructure and required human resources in various fields of ICT.
- ✓ People with special needs must have access to ICT facilities as per their special needs.
- ✓ ICT would lead to e-learning which would result in having any time and anywhere access to learning to make it student driven.
- ✓ Capacity building measures should be formulated on the need-based system for various stakeholders to sharpen the existing skills and competencies of existing manpower to perform their job descriptions effectively and efficiently.
- ✓ Periodic general awareness, training, and developmental programs are conducted for various levels of employees at Government as well as private sector irrespective of their size/nature of ownership and control.
- ✓ Refresher/ Orientation courses must be regularly conducted for all the stakeholders who are involved in imparting training and knowledge in the area of ICT.
- ✓ The services of the subject experts with proven track records must be utilized in preparing e-content that will be available to the masses.
- ✓ State of the art repositories must be in place in institutions of Higher Learning so that content is available on demand.
- ✓ Possibilities of having public/private partnership must be explored for achieving the objectives of ICT.

5.0 Conclusions

The education plays a key role within society in the process of social change. Therefore, the target in the societies has changed to a mass system of education. Our vision of teachers and learners immersed in a network of rich learning objects that are continually enriched and enhanced by the participants is achievable with anticipated advances in information technology and learning sciences. As a result of these efforts, we will be closer in achieving such goals as having transformed the country into a knowledge society. These new benefits will be facilitated by geometric advances in semiconductor and magnetic storage, as well as in electronic and optical communications. The international education and training market are a highly competitive industry. Education providers are increasingly investing in innovative and sophisticated marketing. The use of communications and information technologies in education development and delivery is vital if our educational institutions are to achieve competitive success in the international market for higher and professional education.

6. REFERENCES

[1] MEHARI, Information risk analysis and management methodology, V3, Concepts and Mechanisms, CLUSIF, October 2004.

[2] Alexander, S. Teaching and Learning on the World Wide Web, Paper presented at AusWeb95, the first Australian World Wide Web Conference.

[3] Mehraj ud Din Dar, Muheet Ahmed Butt, Majid Zaman Baba, "Challenges in Educational System: Scope for E-Support", J & K Science Congress University of Kashmir. 25-27 July 2006

[4] Austin, P. & Vaughan, C. 1997, Edith Cowan University Web Enrolment System ECUWES, Paper presented at AusWeb97, the Third Australian World Wide Web Conference, 5–9 July 1997.

[5] William Aspray, Mary J. Irwin "Grand Research Challenges", June 2002, Airlie House in Virginia.

[6] Battacharya & Saxena, "Information Security" World Comp 2007, June 25-28, 2007, Las Vegas, Nevada, USA, ISBN1-60132-046-9.

ISBN: 1-60132-441-3, CSREA Press ©

SESSION

SOCIAL NETWORKS AND SOCIAL MEDIA

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

Revealing Roles of Actors in Clandestine Social Networks

A.Kiruthiga[1] and Dr S.Bose[2]
Department of Computer Science and Engineering, College of Engineering Guindy,
Anna University, Chennai- 600 025, Tamilnadu,India.
E-mail Address1: kiruthiga312@gmail.com
E-mail Address2: sbs@annauniv.edu
*Corresponding author: Dr S.Bose

Abstract - *Role played by an actor is a subjective characterization of the part it plays in the network structure. Actors in Clandestine Social Networks play variety of roles to accomplish their goals. Knowing the role of a terrorist is important in Clandestine Social Network Analysis to device successful strategies to defeat them. Social Network Analysis metrics were extensively used to tag the roles of terrorists in their Clandestine Social Networks. Centrality measures like degree, closeness and betweeness can be used to assign roles to individual nodes in a network; none of these takes the community structure into account for role assignment problem. Community based metric which is inclined towards the cliques in the network as the basis for community structure failed to reveal the roles which plays the intermediate roles in between the communities. In Clandestine Social Networks a single node usually plays different roles, but in the literature, the nodes were assigned with only one role, to overcome the problem of being biased towards either community oriented or community free single role assignment, the process of revealing community based and community free role for each node in the Clandestine Social Network and prioritization of roles are proposed in this research work.*

.***Keywords:*** *role analysis, terrorist network, position analysis, Clandestine Social Network.*

1. Introduction

The word terrorism comes from the Latin word "terror". Terror comes from the Latin word "terrore", which means "tremble" or "frighten". When the French suffix "isme" which means "to practice" is added to "terrere" , it becomes terrorism , which in turn means "practicing the trembling" or "causing the frightening" . Trembling and frightening are synonym for terror. The word terror is over 2,100 years old [1].

A social network is a social structure which is made up of a set of individuals or organizations as actors and a set of ties between these actors. Social Network Analysis (SNA) is the analysis of social networks. SNA views social relationships in terms of network theory, consisting of nodes which represent individual actors within the network and ties which represent relationships between the individuals.

Clandestine (terrorist) organizations are well suited to be analyzed using SNA, as they consist of networks of individuals. Specifically, SNA can be used to analyze, understand clandestine networks and frame more effective anti-terrorism strategies [2].

A node role is a subjective characterization of the part it plays in a network structure. Role analysis is the process of identifying what role a node plays in network either with respect to the community or as a whole network. Each node is assigned with appropriate roles. It also includes the expansion and refinement of role signature and identification of important roles.

The remaining structure of the paper is arranged to explain the proposed framework of Role Revealer model. The following section presents the literature review on role analysis in Clandestine Social Networks (CSN). The later sections present the proposed Role Revealer model and discusses the results of the same. Finally, this paper concludes by mentioning the future enhancements that could improvise the proposed model.

2. Literature Review

Aparna Basu used Betweeness centrality to identify the organization which played brokerage role in a group of clandestine organizations operated together [3]. Jeffy Victoroff has classified the roles within the terrorist hierarchy as sponsor, leader, executive, committee, middle-management and follower. No algorithm proposed for discovering the above said roles in a given CSN [4]. Nasrullah Memon and Henrik proposed a method to select important nodes in an

network by computing the change in the efficiency of the network and they named them as leaders and rest of the nodes as followers [5]. Borgatti selected key players in a given network as KPP-POS and KPP-NEG, based on their purpose of them in the given network as information diffusers and network fragmentors respectively [6].

Carlo Morselli used degree and betweeness centrality to tag leaders and brokers in a network respectively [7]. Ala Berzinji proposed an algorithm to tag finance manager in a given CSN using degree, betweeness and closeness centrality [8]. Tim Minor used the degree centrality to tag hubs in the network[9]. Richard M.Medina used centrality measures to tag leaders and brokers [10] and Olivier Walther tagged brokerage role in a network[11]. Based on the above literature investigation, the roles of actors in CSN was analyzed only based on centrality measures and only leader and broker roles are majorly taken on common basis. The objective of this research work is to identify multiple roles of actors in CSN bassed on community and non-community.

3. Role Revealer Model

3.1 Overall System Design

The Role Revealer model takes social network in the form of adjacency matrix of CSN instance as input. Role assignment based on both on community and community free is performed firstly, and then secondly role prioritization is used for prioritizing roles based on the standard priorities available as well as relative weightage. Figure 1 represents the schematic form of Role Revealer model.

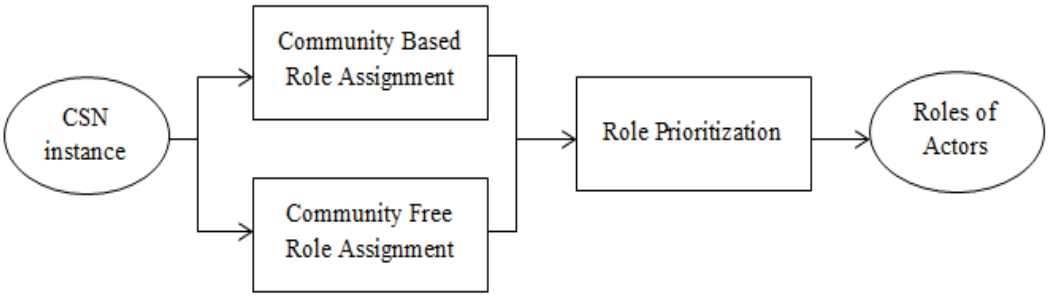


Figure1. Role Revealer Model

3.2 Community Based Role Assignment

The community oriented roles are assigned in two phase as in Figure 2. In the first phase, the network is analyzed and neighbor of each node is computed. Based on each nodes neighborhood the relations like Strongly Connected (SC) [12] and Common Intermediate Node (CIN) [12] are computed. Based on definitions of each roles which are applied on SC and CIN, the community oriented roles are assigned. In the second phase the nodes are grouped into community based on Smart Local Moving algorithm [13]. Based on the community, the link-community probabilities [14] are computed. Using these probabilities the node-community measure [14] is computed .Then the community oriented roles are assigned based on both node- community measure and degree of each node.

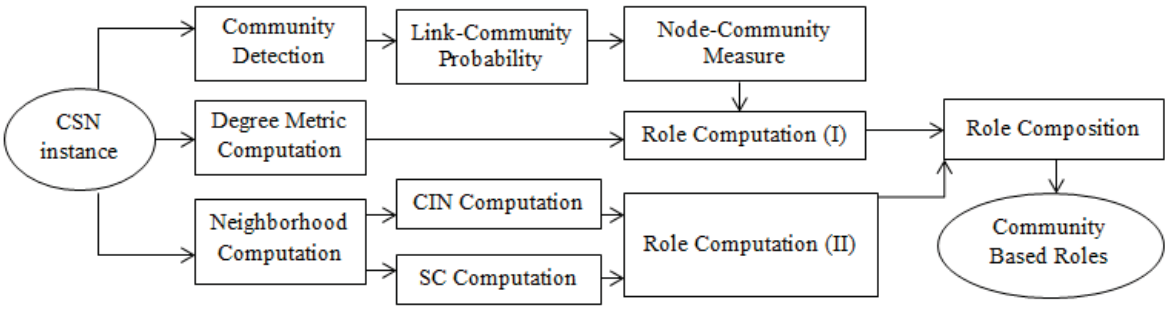


Figure 2. Community Based Role Assignment

The respective algorithm for community based role assignment which is shown in Figure 2 is given below.

CommunityBasedRoleAssigment (CSN G)
{

```
  for each node k
  {
    neighbour of k = getNeighbors (k)
    communityFormer(k)
    computeP(g)
    computeQ(g)
    computeTau(k)
    roleComputation1(k)
    if (k's role not assigned)
    {
      roleComputation2(k)
    }
  }
}
```

```
roleComputation1(node k)
{
  if(assignBridge(k)) return
  if(assignGateway(k)) return
  if(assignHub(k))
}
```

```
roleComputation2(node k)
{
  if(assignBigFish(k)) return
  if(assignAmbassador(k)) return
  if(assignBridge2(k)) return
  if(assignLoner2(k))
}
```

The nodes are analyzed for its neighbors and based on definitions of CIN and SC the first set of roles are assigned. Then the node-community measure is computed by grouping the nodes to community. Based on the node- community measure and degree metric, the second set of roles are assigned. The functions used in the above algorithm are explained below.

getNeighbors(node v) function returns the neighbors of the node v. communityFormer(CSN g) function groups the nodes as communities of the given CSN instance g. Link-Community probability is computed using computeP(CSN g), computeQ(CSN g) and computeTau(node k) functions. computeP(CSN g) function Computes the probability for two linked nodes that are in same communities. In Equation 1 a complete node pair [15] is one where the linked nodes belong to the same community.

$$p=\frac{complete\ node\ pairs}{total\ linked\ node\ pairs} \quad (1)$$

computeQ(CSN g) function computes the probability for two non-linked nodes that are in different communities. In Equation 2 pure node pairs [15] are non-linked nodes that do not appear together in any community.

$$q=\frac{Pure\ node\ pairs}{Total\ non-linked\ node\ pairs} \quad (2)$$

computeTau(node k) function computes the elementary value for community metric of each node. Equation 3 is used to compute the community metric.

$$\tau_u(v_i)=\frac{1}{1+\sum_{v_j\in N(u)} I(v_i,v_j).p+\bar{I}(v_i,v_j).(1-q)} \quad (3)$$

Node-Community measure is computed by the function namely computeCommunityMetric(v) which normalizes the community metric using rawComm as shown in Equation 4 &5.

$$rawComm(v)=\sum_{v\in N(u)}\tau_u(v) \quad (4)$$

$$communityMetric(v)=\frac{rawComm(v)-minrawComm}{maxrawComm-minrawComm} \quad (5)$$

isCIN(node a,node b) function returns true if nodes a and b have only one common neighbor node. The neighbors of both a and b are computed and intersection operation is performed. If the result length is 1, boolean value true is returned else false is returned. isSC(node a,node b) function returns true if the nodes a,b are strongly connected.That is they have two or more common neighbors. The intersection of neighbors of both nodes a and b is found and the length of result is calculated. If the length is greater than or equal to 2, boolean true is returned else boolean false is returned. assignBridge(v) function assigns Bridge [12] role as per the following definition.

For all pair of nodes x and y which belongs to neighbor of v other than v, they should be in CIN and they should not be loner. assignGateway(v) function assigns Gateway [12] role to nodes which satisfies the following two conditions. There exists two different nodes x and y other than v and neighbors of v are in SC.Second, it has another neighbor that is not a loner and does not share any common neighbor except v with v's other neighbors.

assignHub(v) : A vertex v is called Hub [12] if there exist w, x, y and z which are neighbors of v and which satisfy the following conditions. w and x are in SC, and y and z are in SC as well. W and y are in CIN, and x and z are in CIN as well. assignLoner(v) functions assigns a node with Loner [12] role if it has exactly two neighbors including v.

computeDegreeMetric(v) : The degree metric of each node is normalized between 0 and 1. assignAmbassador(v) method assigns the Ambassador [14] role for a node if it has higher normalized community metric and higher normalized degree metric. assignBigFish(v) method assigns the Big Fish [14] role to the nodes which have relatively lower community metric and higher degree metric. Nodes with higher community metric and lower degree metric are assigned with Bridge role by assignBridge(v) method. Loner [14] roles are assigned to nodes which have lower community metric and lower degree metric by assignLoner2(v) method.

3.3 Community Free Role Assignment

Community Free Role Assignment as in Figure 3 assigns the roles to each node in the network with global aspect. In this module the roles are categorized as degree rich roles, degree poor roles and connection oriented roles. Roles such as Hub and Influencer [16] are under degree rich category. In order to find Hub, degree metric is needed. To find Influencer, both degree and betweeness metrics are needed. In degree poor roles, Fringe [16] and Loner are there. To find Loner, degree metric is needed. For finding Fringe, additionally closeness metric is required. Bridge and Gateway are connection oriented roles. BASSET gatewayness algorithm [17] is used to find them.

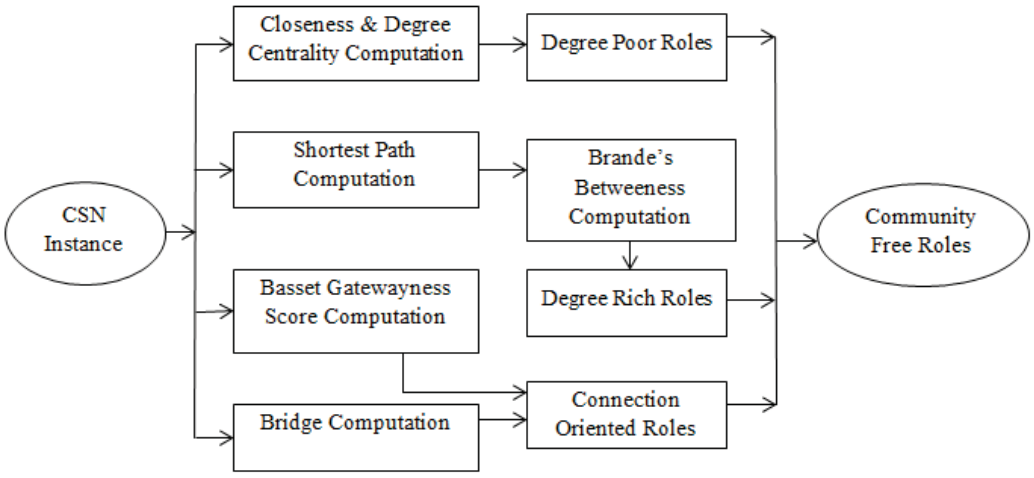


Figure 3. Community Free Roles Assignment

```
---------------------------------------------------------------------------
CommunityFreeRoleAssigment(CSN g)
{
  for each ndoe n
    degree = computeNeighbours(n)
    computeShortestPath(g)
    computeBetweeness(g)
    computeCloseness(g)
    computeGateway(g)
    computeBridge(g)
    assignRole(n)
}
---------------------------------------------------------------------------
computeGateway(CSN g)
{
  initialize i to be empty.
  compute the proximity score r(s,t) from the source node s to
   the target node t.
  find argmax
  for j =2 to k do
  for i =1 to n do
  let J=I U i
  compute v(i)
  end for
  if max(i) <=r(s,t) then
  find i0 =argmax(i); add i0 to I;
  else
  bresk;
  end if
  end for
  return I
}
---------------------------------------------------------------------------
assignRole(n)
{
  if(assignHub (n)) return
  if(assignLoner(n)) retun
  if(assignFringe(n)) return
  if(assignInflencer(n)) return
  if(assignBridge(n)) return
  if(assignGateway(n)) return
}
---------------------------------------------------------------------------
```

The functions used in the above algorithm are explained below. computeNeighbours() is used to calculate respective degree metrics of each node. computeShortestpath() is used to calculate the shortest distance between the each node with each other node. computeCloseness() is used to find the shortest distance between the most centrality node with each other nodes. computeBridge() is used to find the bridges in the network. computeBetweeness() function calculates betweeness by using an algorithm called Brandes algorithm [18] it was devised by a German scientist Ulrik Brandes. The shortest path is calculated using Djikstra's algorithm [19].

After finding all the metrics needed, roles for each and every node can be assigned. assignHub() assign the role hub to the node based on degree metric. assignLoner() assign the role loner to the node based on the degree metric. assignFringe() assign the role fringe to the node based on the closeness centrality. assignInfluencer() assign the role influencer to the

node based on the betweeness and degree metric. assignGateway() assign the role gateway to the node based on the gateway-ness score. assignBridge() assign the role bridge to the node.

3.4 Role Prioritization

Role prioritization Module which is represented in the Figure 4 takes both community based and community free roles as input and prioritizes them. Since each node is assigned with both community based and community free roles, Role Aggregator groups them into a single role list which is then used for prioritizing based on their importance. Top Role Revealer reveals the top(important) role for a node by comparing both community based and community free roles. Roles for each node is assigned based on their priority. Priorities to each roles is being set either relative to the network or by using standard priorities available. Prioritized roles for each node is then used for role denotation and justification.

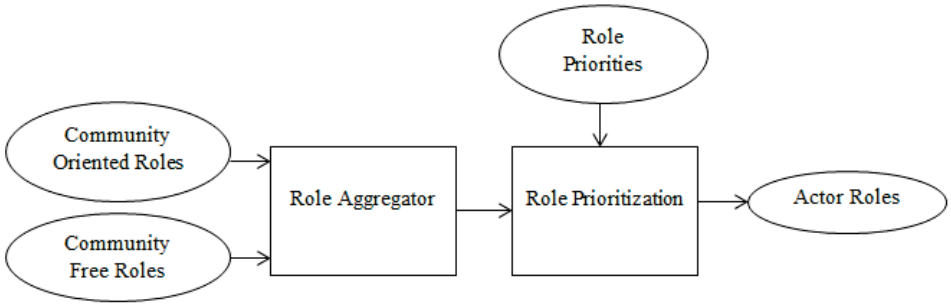


Figure 4. Role Prioritization

```
---------------------------------------------------------------------------
RolePrioritization (CommBasedRoles,commFreeRoles)
{
  //Aggregate both commFreeRoles and commBasedRoles

  roleList=roleAggregate()

  //Reveal Top Role for each node

  for(node:n)
  {
    if(roleList(i)<roleList(i+1))
      reveal(role(i+1))
    else
      reveal(role(i))
}
---------------------------------------------------------------------------
```

Role prioritization module aggregates both community based and community free roles and reveals the top role based on the priorities. roleAggregate() function Aggregates both community based and community free roles into single role list as per the definition based on the role assignment algorithm. These aggregated roles are used for prioritizing roles based on the definition.

reveal(role(i)) function works in the following way; First the priority for a role is set either by relative to the network or by standard priorities. Relative importance to a role is assigned by calculating the weightage for a particular with respect to the network. Standard priorities are set based on the references. Each node is assigned with high priority role.

4. Results

Role distribution both communities based and community free were analyzed in CSN datasets. Sample results were given in Figure 5 and 6. The datasets used for analysis were namely Bali bombing network [20], Bojinka network[21] and Madrid bombing network [21].

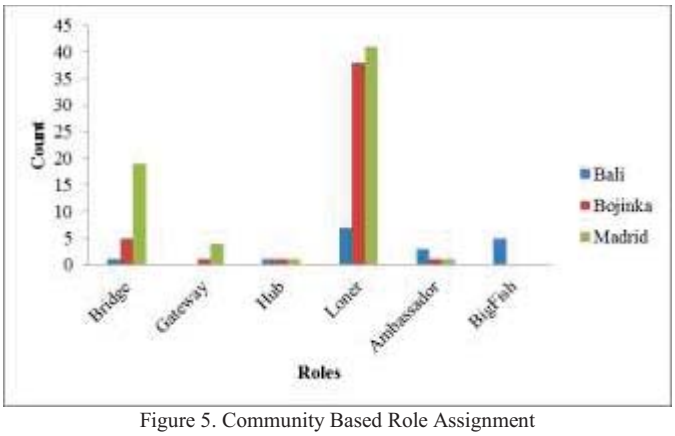


Figure 5. Community Based Role Assignment

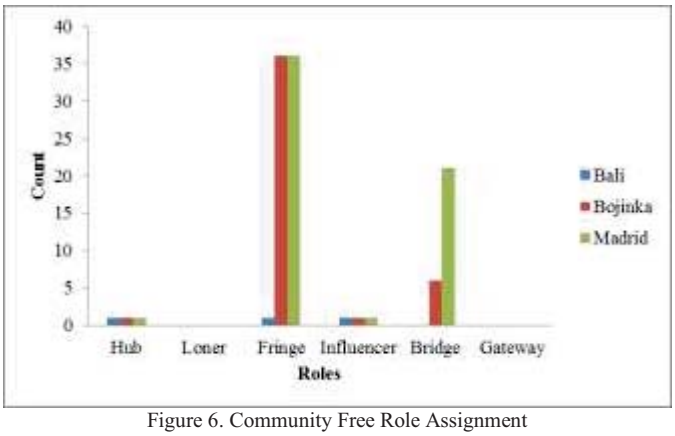


Figure 6. Community Free Role Assignment

5. Conclusion

The model presented in this paper reveals roles of actors in CSN based on community and community-free. Multiple roles are assigned for each actor which significantly increases the understanding of the CSN which helps the counter-terrorism agencies to frame effective strategies. Validation of the roles played by each actor in a real-time should be done to improvise the proposed model.

6. References

1. Burgess, Mark. "A brief history of terrorism" *Center for Defense Information*, 2003.
2. Ressler, Steve. "Social network analysis as an approach to combat terrorism: Past, present, and future research"; Homeland Security Affairs, 2, 2, 2006.
3. Basu, Aparna. "Social network analysis of terrorist organizations in India"; North American Association for Computational Social and Organizational Science (NAACSOS) Conference, 26-28, 2005.
4. Victoroff, Jeff. "The mind of the terrorist A review and critique of psychological approaches "; Journal of Conflict resolution, 49, 1, 3-42, 2005.
5. Memon, Nasrullah, and Henrik Legind Larsen. "Structural Analysis and Destabilizing Terrorist Networks"; DMIN, 296-302. 2006.
6. Borgatti, Stephen P. "Identifying sets of key players in a social network "; Computational & Mathematical Organization Theory, 12, 1, 21-34, 2006.
7. Morselli, Carlo. "Assessing vulnerable and strategic positions in a criminal network "; Journal of Contemporary Criminal Justice, 26, 4, 382-392, 2010.
8. Berzinji, Ala, Lisa Kaati, and Ahmed Rezine. "Detecting key players in terrorist network "; Intelligence and Security Informatics Conference (EISIC), 297-302, 2012.
9. Minor, Tim. "Attacking the Nodes of Terrorist Networks"; Global Security Studies 3, 2, 1-12, 2012.
10. Medina, Richard M. "Social network analysis: a case study of the Islamist terrorist network"; Security Journal, 27, 1, 97-121, 2014.
11. Olivier, w. a. l. t. h. e. r., and c. h. r. i. s. t. o. p. o. u. l. o. s. Dimitris." A social network analysis of islamic terrorism and the malian rebellion", 2012-38, 2012.
12. Chou, Bin-Hui, and Einoshin Suzuki. "Discovering communityoriented roles of nodes in a social network"; Data warehousing and knowledge discovery, 3, 52–64, 2010.
13. Waltman, Ludo, and Nees Jan van Eck. "A smart local moving algorithm for large-scale modularity-based community detection"; The European Physical Journal, 86, 11, 108–113, 2013.
14. Scripps, Jerry, Pang-Ning Tan, and Abdol-Hossein Esfahanian. "Node roles and community structure in networks" ;Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 26–35, 2007.
15. Scripps, Jerry, Pang-Ning Tan, and A-H. Esfahanian, "Exploration of link structure and community-based node roles in network analysis", Data Mining,Seventh IEEE International Conference, 649–654, 2007.
16. Chen, Duan-Bing, Hui Gao, Linyuan Lu, and Tao Zhou. "Identifying influential nodes in large-scale directed networks: The role of clustering", 2013.
17. Tong, Hanghang, Spiros Papadimitriou, Christos Faloutsos, S. Yu Philip, and Tina Eliassi-Rad. "Gateway finder in large graphs: problem definitions and fast solutions"; Information retrieval, 15, 3-4, 391–411, 2012.
18. Brandes and Ulrik. "A faster algorithm for betweenness centrality"; Journal of Mathematical Sociology, 25, 2, 163– 177, 2001.
19. Zhang, Fuhao, Ageng Qiu, and Qingyuan Li. "Improve on dijkstra shortest path algorithm for huge data"; Chinese academy of surveying and mapping, 2005.
20. Koschade, S. "A social network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence"; Studies in Conflict & Terrorism, 29, 6, 559-575, 2006.
21. Memon, N., Larsen, H. L., Hicks, D. L., & Harkiolakis, N. "Retracted: Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies"; Intelligence and Security Informatics, 477-489, Springer Berlin Heidelberg, 2008.

Computing Social Communities in Asymmetric Social Media like Twitter

Gurpreet Singh Bawa
Gurpreet.singh.bawa@accenture.com

Abstract - *An asymmetric Social Media source like Twitter is somewhat like a "Citizen's Band" radio in that it allows people to follow others irrespective of reciprocity: because someone follows me I don't necessarily follow them. It is more challenging to define social network in this form of media since, unlike a 'friend' linkage, a link in this network does not conform to the mutual reciprocity that is normally expected.*

In this paper, we demonstrate a social network calculation approach that works with both symmetric social media sources like Facebook as well as asymmetric sources like Twitter. We demonstrate the differences between the networks and how the revealed structure can be used to interpret topics of conversation, group influence and other important indicators of social activity.

Keywords: Social Communities, Twitter, Social Media, Network Analysis

1 Introduction

Social network analysis (SNA) is a set of theories, tools, and processes for understanding the relationships and structures of a network. The "nodes" of a network are the people and the "links" are the relationships between people. Nodes are also used to represent events, ideas, objects, or other things. SNA practitioners collect network data, analyze the data (e.g., with special purpose SNA software), and often produce maps or pictures that display the patterns of connections between the nodes of the network. The maps in this article were created using SNA computer programs by Borgatti (2002) and Brandes and Wagner (2004).

Many mathematical techniques are available to measure networks (Wasserman & Faust, 1994); below we highlight a few particularly relevant. We will also demonstrate how to use these metrics to understand and evaluate specific networks.

1.1 Bonding and bridging

Bonding and bridging are two different kinds of connectivity. Bonding denotes connections in a tightly knit group. Bridging denotes connections to diverse others. See Fig. 1 for an illustration. These terms are commonly used in the social capital literature (Putnam, 2001). In the SNA literature, bonding and bridging are often called "closure" and "brokerage" respectively (Burt, 2005); also, "strong ties" and "weak ties" are important related SNA concepts that we incorporate into our bonding–bridging usage (Granovetter, 1983). Analyzing network data to measure bonding and bridging helps to predict important outcomes such as efficiency and innovation: bonding indicates a sense of trusted community where interactions are familiar and efficient; bridging indicates access to new resources and opportunity for innovation and profit (Burt, 2005).

1.2 Clusters

A cluster is a tightly knit, highly bonded, subgroup. Identifying clusters is one of the most important applications of SNA, because it illuminates important previously unrecognized subgroups. Clusters can be displayed visually with a network map, as shown by the three highlighted clusters in Fig. 1. Algorithms that identify clusters measure variations in density and links per node. Density and links per node and core and periphery structures are fundamental network metrics described below.

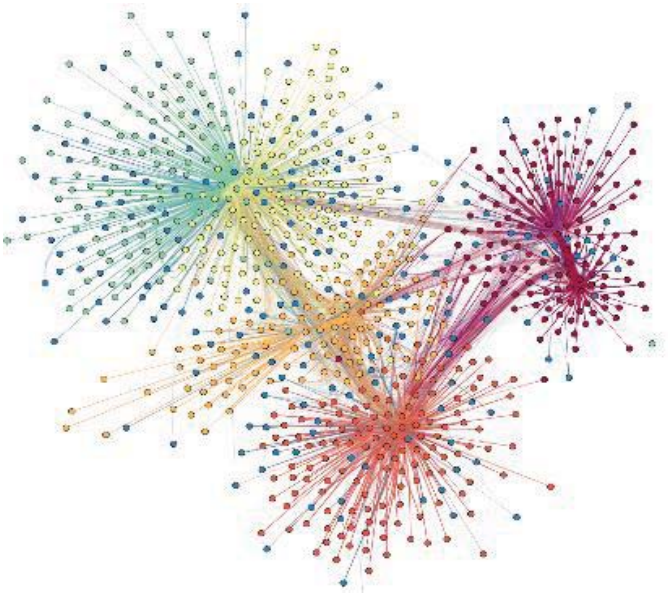


Fig. 1. Bonding, bridging, and clusters.

1.3 Core and periphery

Many networks feature a core/periphery structure. The core is a dominant central cluster, while the periphery has relatively few connections (Borgatti & Everett, 1999). See Fig. 2 for an illustration. Nodes at edges are periphery, while nodes in the center are core.

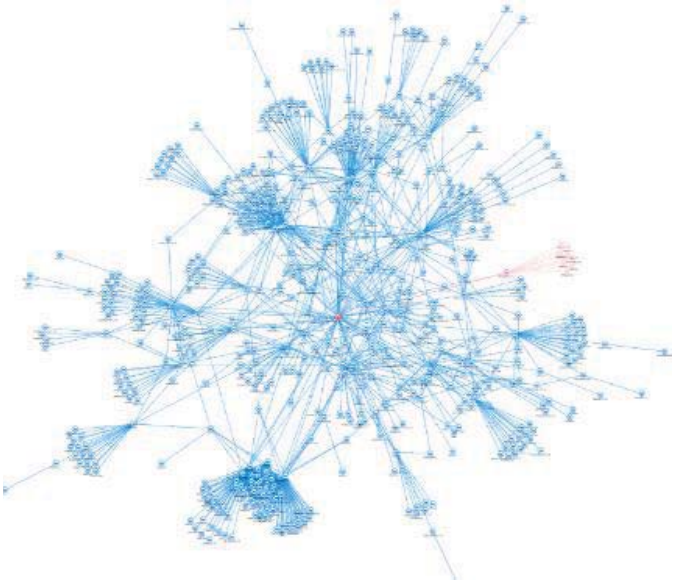


Fig. 2. 1.3 Core and periphery Structure

1.4 Directed and undirected links

Links can be undirected (e.g., "shares information with") or directed (e.g., "seeks advice from"). Directed links can be one-way or two-way. Social network analysis addresses both undirected and directed networks.

1.5 Density and links per node

Density is the number of links that exist in a network divided by the maximum possible number of links that could exist in the network. All of the social network analysis metrics in this paper assume that the numbers of nodes and links that exist in a network are known; we use N to refer to the number of nodes and M to refer to the number of links. The maximum possible number of links in a network depends on N and on whether the network is undirected or directed. For an undirected network, the maximum possible number of links is $N(N-1)/2$; for a directed network it is $N(N-1)$.

Density helps to define clusters. A cluster is a local region in a network with relatively high density and relatively few links to other clusters. Formal mathematical definitions of clusters and algorithms for finding clusters are reviewed by Brandes and Erlebach (2005). Links per node is the total number of links divided by the total number of nodes in the network.

1.6 Bridger's and Between-ness centrality

Bridger's are individuals in a network who have connections to different clusters. Finding bridgers is the "ip side of finding clusters. Bridgers can be highlighted visually just as clusters can; Fig. 1 illustrates a bridger. Bridgers in a leadership network provide valuable opportunities for innovation, growth, and impact because they have access to perspectives, ideas, and networks that are otherwise unknown to most network members. Bridgers are easy to overlook because the significance of their ties is not visible by counting the number of ties. Finding bridgers is an important application of SNA in leadership networks. Bridgers often make good key informants during an evaluation because of their access and knowledge of the larger network.

Finding bridgers in a network is typically done with the calculation called betweenness centrality (Freeman, 1979). This calculation indicates how often one individual is likely to be an important relay point between other network members. Another metric used to find bridgers is network constraint (Burt, 2004, 2005). An individual's network constraint measures the extent to which he links to others that are already linked to each other. Low network constraint means that an individual has links to others who are not already linked to each other. High betweenness centrality and low network constraint both indicate bridging.

1.7 Hubs and Indegree centrality

Hubs are individuals in a network with the most influence. Whether hubs bridge across clusters or bond within a cluster (or some combination), they are highly sought-after by other network members. Hubs of influence in a network are best measured using directed links. Given a network of directed relationships, indegree centrality (or just "indegree") counts how many relationships point towards an individual; this provides a simple measure of influence (Freeman, 1979). More advanced influence metrics build on indegree and consider not just how many others seek the advice of a particular person, but also how influential those other advice-seekers are. A person whose advice is sought by someone who is highly influential may have a higher influence score than one whose advice is sought by many non-influencers. Bonacich and Lloyd (2001) overview several advanced influence metrics and explain how most of them compute nearly the same thing. In most cases, we recommend using indegree, because it communicates the basic point without unnecessary complications.

1.8 Structural equivalence

Amazon.com made structural equivalence famous as the calculation behind its recommendations: "People who bought books A and B also bought books C and D." This Amazon.com example considers both people and books as members of a single network. Links in this network join people to the books they have purchased. People who buy mostly the same books have high structural equivalence; people who buy mostly different books have low structural equivalence.

2 Social Network Analytics

Physical and virtual threads are encoded as links (or edges) between nodes (or vertices); hence the associated

conversations may be analyzed by a range of social network analytic techniques. Of particular importance is the calculation of the sub-nets that characterize a given collection of text documents. Number of other metrics that are used in sub-net and individual node characterization, for example:

- Centrality
- Betweeness
- Influence

Since these metrics are produced on a standard scale they can be summed and averaged across various sub-nets (and hence can be used to compare sub-net attributes).

The community detection of sub-nets in the collection of text documents that are conceived of as the "host" for the various conversations that are monitored in our approach is a primary feature and capability of this conversational approach to text analytics. It is axiomatic that vocabulary and term usage is socially determined since language is a tool for social communication. While the word "snow" might suffice in a general conversation, it is likely that a group of skiers will use such terms as "corn", "sugar", "hard pack", "boilerplate" and so on the describe the many varieties of snow (yet the generic term "snow" may be completely missing from the conversation).

This core capability of social context analytics is missing from the major forms of natural language processing and text topic derivation that are in use today. The ability to situate a conversation in a social context goes a long way towards the precise allocation of meaning to phrases like "bank", for example, that could be referring to a "river bank" or a "savings bank". While it is normal for most text analytic systems to explore the associated vocabulary of the embedded term – to find associated instances of "water" or "financial institutions", it is unknown to characterize the sub-net as "water resource related" or "financial institution related", for example. This example is meant to illustrate the differences between our method and the normal procedure and will suggest how our method is at least different, and possibly superior.

3 Calculate Network descriptors

Each sub-net is uniquely defined by the input fields that are used to create it and the field values that are used to form the branches of a decision tree that defines the hierarchical branch attributes of the sub-net. The hierarchical branch attributes are formed by applying 3 types of sub-net discriminators:

(1) individual, (2) social and (3) operational

Textual expression emerges from a process that is visualized in Figure 4.

The middle part of Figure 4 illustrates the trigger event – in this case the use of a particular expressive term – that results from the top-down flow of the environment and circumstances of the conversation. The environmental context is translated through the conceptual apparatus (neurochemistry, psychology, physiology) of the message originator and results in the generation of a given expression, as shown in the bottom part of Figure 4.

We use the sub-net descriptors indicated here to operationalize the environment and circumstances that affect the triggering expression-generating event. These sub-net discriminators are as follows:

1) Individual characteristics include:

- Age
- Gender
- Education
- Marital status
- Interests, Affiliations and memberships
- Psychological (e.g. behavioral profile)

2) Social organization characteristics include:

- Geographic and temporal location
- Social role: Leader, follower, marginal
- Social Influence
- Community size, density, dispersion
- Community character (for example, friends, family, business associates, social, recreational and spiritual groups)

The messages that are exchanged in the conversation also have operational characteristics that can impact the sense characterization.

3) Operational characteristics include:

- Message recency
- Message frequency
- Conversation acceleration rate
- Message mood state
- Type of exchange: personal, professional

The specific form of the sub-net is calculated using the predictive modeling capability. This capability shows the relationship between a target value (the Topic) and various input values (individual, social and operational inputs). The decision tree is formed by searching for important relationships between the topics that have been extracted at the top-level text corpus and the three sets of field values that have been calculated as sub-net discriminators.

A causal sequence is implied in the unfolding of the sub-net characteristics. As shown in Figure 3, an element of the conversation is filtered through the participants' individual characteristics, social characteristics and message characteristics, in that order. (Clearly, the individual background will influence social choices and this context will

further be influenced by the operational features of the messages that are exchanged in the conversation.)

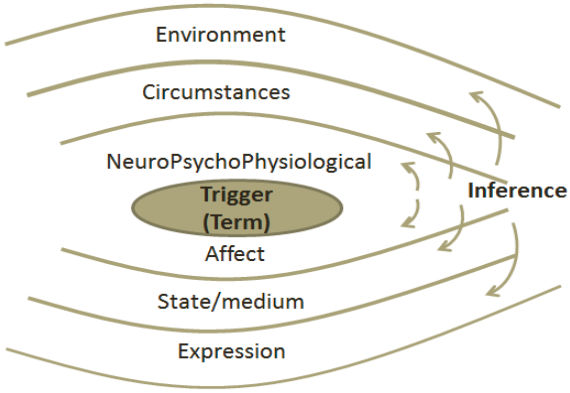


Fig 4: Environmental Context of Conversational Expression

4 References

[1] Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. Social Networks, 23(3), 191−201.

[2] Borgatti, S. P. (2002). NetDraw: Graph visualization software. Harvard: Analytic Technologies.

[3] Borgatti, S. P. (2005). Centrality and network "ow. Social Networks, 27(1), 55−71.

[4] Borgatti, S. P., Carley, K., & Krackhardt, D. (2006). Robustness of centrality measures under conditions of imperfect data. Social Networks, 28, 124−136.

[5] Borgatti, S. P., & Cross, R. (2003). A relational view of information seeking and learning in social networks. Management Science, 49(4), 432−445.

[6] Brandes, U., & Wagner, D. (2004). Visone: Analysis and visualization of social networks. In Michael Jünger, & Petra Mutzel (Eds.), Graph Drawing Software(pp. 321−340). New York: Springer-Verlag

[7] Evans, P., & Wolf, B. (2005). Collaboration rules. Harvard Business Review, 83(7), 96−104.

[8] Feller, J., Fitzgerald, B., Hissam, S., & Lakhani, K. R. (Eds.). (2005). Perspectives on free and open source software. Cambridge: MIT Press.

[9] Freeman, L. (1979). Centrality in networks: I. Conceptual clari!cation. Social Networks, 1, 215−239.

[10] Friedman, M. (2005). Trying hard is not good enough: How to produce measurable improvements for customers and communities. Victoria, B.C.: Trafford Publishing.

[11] Gajda, R., & Koliba, C. (2007). Evaluating the imperative of intra-organizational collaboration: A school improvement perspective. American Journal of Evaluation, 28(1), 26−44.

[12] Gauthier, A. (2006). Developing collective leadership: Partnering in multi-stakeholder contexts. In W. Link, T. Carral, & M. Gerzon (Eds.), Leadership is Global.

[13] Gower, J. C. (1971). A general coef!cient of similarity and some of its properties. Biometrics, 4(27), 857−871.

[14] Granovetter, M. S. (1983). The strength of the weak tie: Revisited. Sociological Theory, 1, 201−233.

[15] Grove, J., Kibel, B., & Haas, T. (2007). EvaluLEAD: An open-systems perspective on evaluating leadership development. In K. Hannum, J. Martineau, & C. Reinelt

[16] (Eds.), Handbook of leadership development evaluation. San Francisco: Jossey-Bass.

[17] Gutierrez, M., Tasse, T., Gutierrez-Mayka, M., & Hagen, G. (2006). Assessment of the Annie E Casey Foundation's Children and Family Fellowship Program.

[18] Unpublished Evaluation.

[19] Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods Retrieved March 20, 2008, from the University of California, Riverside Web site:

[20] http://www.faculty.ucr.edu/~hanneman/

[21] Kilduff, M., & Tsai, W. (2003). Social networks and organizations. London: Sage.

[22] Krackhardt, D., & Hanson, J. (1993). Informal networks: The company behind the chart. Harvard Business Review, 71(4), 104−111.

[23] Krebs, V., & Holley, J. (2002). Building smart communities through network weaving Retrieved April 30, 2007 from Orgnet.com: http://www.orgnet.com/

[24] BuildingNetworks.pdf

[25] Kunkel, P. (2005). Collective leadership—A pathway to collective intelligence. Collective Leadership Institute Retrieved on October 5, 2006 at http://www.

SESSION

KNOWLEDGE EXTRACTION AND ENGINEERING, WEB AND MOBILE COMPUTING, ONTOLOGY MAPPING, AND APPLICATIONS

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

KNOWLEDGE FRAMEWORK ON THE EXECUTION OF COMPLEX PROJECTS

-THE DEVELOPMENT OF A FUNCTIONAL FRAMEWORK USING A SYSTEMS APPROACH-

J.Wiskerke, H.Veeke, J. Pruijn, C. Groen, H. Hopman
jeroen.wiskerke@damen.com, H.P.M.Veeke@tudelft.nl, J.F.J.Pruyn@tudelft.nl,
keesjan.groen@damen.com, J.J.Hopman@tudelft.nl

Abstract

This paper describes a functional framework that has been developed to realize a steady execution of a variety of complex projects. It provides knowledge to successfully intervene in a currently problematic situation. To achieve this, theory and practice have been equal constituents. The development is based on the Delft Systems Approach and the experience is gained from three executed projects at ship yards. The framework has been tested on project data and expert reviews, judging fit, relevance, workability and modifiability of the framework.

Keywords: Knowledge Management, Systems Approach, Framework, Complexity

Introduction

The relation between the multidisciplinary nature of a design or maintenance and repair (M&R) project and the mainly mono-disciplinary participants of the project team, is a major problem during execution. It leads to a clear misfit between intentions and perceptions of the project result. As long as the complexity of a project is relatively small, the misfit can be solved during execution. But with complex projects - as occurring in innovative design or non-recurring M&R - the misfit has consequences like unexpected results or costs exceeding the budget. In a former paper [Veeke, Lodewijks et al, 2006] conceptual modelling was proposed as a generic interdisciplinary activity rather than a domain-specific one, in order to bridge the gap. It was shown that the Delft Systems Approach [Veeke, Ottjes, et al, 2008] offers the basic models to describe any innovative design in a functional/conceptual way.
In this paper, the same approach is used for non-recurring M&R projects at shipyards. After analyzing three complex projects, executed at three different conversion yards, it appeared that the execution of complex projects requires not only a similar way to describe the project activities but also a different project organization.
Today's practices do not fit the requirements for these complex projects. This will result in a decrease of efficiency and introduce the need for a change in the current situation [Tushman, Nadler,1978]. Current project executions, often with a taken-for-granted nature as described in [Nelson, Winter, 1982], are still common practice at ship yards. They are executed as "normal" M&R projects and are enforced by local, historical influences. Suboptimal project executions are rooted in the past, and originated from historical traditions [Levering, Ligthart et al, 2012].
The misfit between current practice and demands shows most clearly in the execution of complex projects as they significantly differ from the normal M&R as described by both Senturk[2008] and Levering et al. [2012]. The execution in complex projects is yard specific, non-repetitive and there is no mutual learning curve between the different yards. Best practices and knowledge are rarely shared.
In order to investigate this problem further, we need to define complex projects at shipyards first. After that we will apply the models of the Delft Systems Approach to these complex projects and discuss the organizational consequences of them for the execution.

Complexity of M&R projects

The two most common types of complexity within projects concern the organizational and technological complexity.

Organizational complexity is caused by the engagement of several separate and diverse organizations for a finite period of time. This leads to a temporary multi-organizational structure to manage a project. The level of complexity depends on the differentiation, the interdependencies and the interaction within the organizational structure [Hall,1979]. Differentiation is, according to [Baccarini, 1996], caused by hierarchical structures and organizational units.
Technical complexity is caused by differentiation and interdependency as well. In this case, differentiation refers to the diversity of tasks.

We define a complex project in this paper as "a project with an estimated value of many millions of euro's, consisting of many interrelated and interdependent parties that work together as one whole. The project is executed under the supervision of a single shipyard, to convert – in the broadest sense – a vessel within a restricted timeframe and budget. It is more demanding than the normal M&R projects with respect to organization, legislation and technology"

A measure for complexity is not available yet within this definition. Complexity is usually expressed by means of cost, duration or numbers of people involved. These criteria however don't correlate well with management complexity [Duncan, 2013].

Originally GAPPS (Global Alliance for Project Performance Standards) developed a framework that categorizes projects based on their management complexity by seven factors. These are based on construction projects, so they are adapted slightly to fit the management of complex M&R projects in shipyards. The factors are:

A. Scope
1. Stability of the overall project context.
2. Number of distinct disciplines, subcontractors, methods or approaches involved.
3. Environmental impact.
4. Financial impact on the organization.

B. Client
5. Client influence on organizational procedures.

C. Organization
6. Experience on the type of work.
7. Impact on planning and capabilities.

D. Subcontractors
8. Cohesion between yard and subcontractors.

To qualify the complexity of a project these factors are rated on a point scale from 1 to 4 (1=Low, 2=Moderate, 3=High, 4=Very High). Of course, this system is somewhat subjective as it depends on the consistency of the assessor(s). Above that, it is subject to a learning curve in the execution of complex projects. But it is still useful in providing an aid to assess the complexity of a project and looking over specific difficulties causing the complexity.
To deal with complexity caused by the factors 2, 3, 4 and 6, the approach as proposed in [Veeke, Lodewijks et al., 2006] would suffice. The other factors influence the complexity of management directly. In an unstable context and/or with little cohesion between yard and subcontractors, it is difficult to keep every participant directed into the right direction. Whenever a client has large influence on the project contents and/or until a late phase during execution, it is difficult to control time and budget of a project.

Case studies

Three executed complex projects are analyzed by means of the available project data. The complexity of the projects was qualified by the factors above and ranges from High to Very High.
The case studies analyzed the problem both from practice and theory. It appears that complex projects in the current situation are executed as enlarged M&R projects, show transient reactive behavior and lack appropriate project control and internal integration. The separate yards work at a certain basis level (S) with respect to their ability to carry out (complex) projects. This level differs from yard to yard and due to the lack of a mutual learning curve, this base is not shared and therefore does not, or only very limited, increase on group level. On the contrary, the introduction of complex projects requires a higher level of project execution which at this moment is not readily available. There is a certain growing gap, Δ, between the current level of project execution and the new demands. The need for a functional framework that provides a singular methodology for the execution of complex projects is identified, describing and covering the growing misfit Δ.
It should allow the tuning to a specific project, while keeping the relations and interdependencies similar.

Problem definition

Most literature only emphasizes correctional control. For example [Kerzner, 11] argues that control is a three-step process: measuring progress towards an objective, evaluating what remains to be done, and initiating the necessary corrective actions to achieve or even exceed the objectives. However, this definition does not include the feasibility check nor the control needed before anything will be done.

[In't Veld, 12] distinguishes four essential conditions that enable a process to be controlled properly:

1. There must be an objective; the expected result needs to be defined.
2. The system should be capable of realizing this objective (feasibility).
3. The behavior of the system should be adjustable
4. The interdependency and relations with the environment should be known.

The first two points express the need of feasible standards and maintaining them, while the other two points express the need of reactive potential in case of disturbances. In DSA these different controls are combined with the operational process into one single "steady-state model"; the model represents a general "function", which is a blueprint for anything happening in an industrial organization (fig. 1).

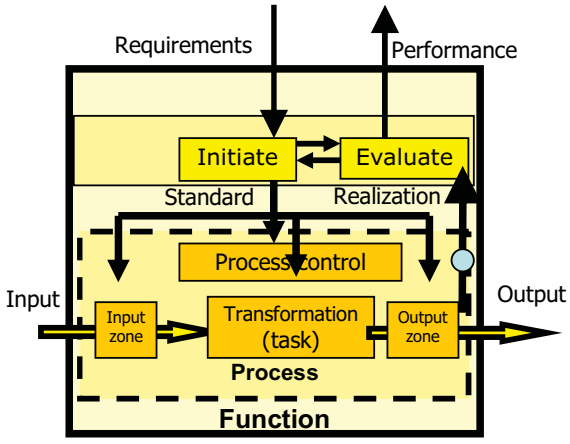


Fig. 1 Steady state model

The steady-state model represents all possible functional activities around one single material flow (it is a so-called "single aspect" model). The activities aim to make a product as required in a controlled repetitive way. When investigating logistic / industrial systems one often gets involved with three different flows:

The material flow: the production
The order flow: nothing is produced without an order
The resource flow: nothing can be produced without resources

In order to get an insight into the relations between these flows DSA contains the PROPER (PROcess PERformance) model, which visualizes the three separate but correlated flows. Each flow can be represented by a steady-state model, while the combination of flows is shown in the PROPER model.
The control functionality should be extended by a coordination control that cannot be present by definition in the steady-state model. Coordination control tunes the different flows to each other.

Both steady-state and PROPER-model are empty with respect to concrete content. So it is necessary to use the practical experiences for developing a functional framework with these models. The PROPER model is adapted to a specific model representing the process of the repeated execution of complex projects in the as-is situation. The cascade of control functions in the higher levels and the different product flows in the lowest level of the model provide insight in the functions and their interactions and interfaces.

On the top level of figure 3 – level V-, the environment is taken as the complement of actors and influences on all levels below. The environment is, but is not limited to: agencies, competitors, the market and governments. It represents everything that has a possible influence on the organization but they are taken into account in the model explicitly as the external requirements.

Level IV in the model is the first step in the cascade of control functions and consists of both the (potential) clients and the highest level of management. By means of acquisition, a client comes from the top level to this lower one with a set of requirements and possibly an Invitation To Tender (ITT) that enters the tender function at level III. Within the organization, at level IV, developments and changes in the requirements from the environment are analyzed and a suitable (long-term) approach is defined by a coordination control.

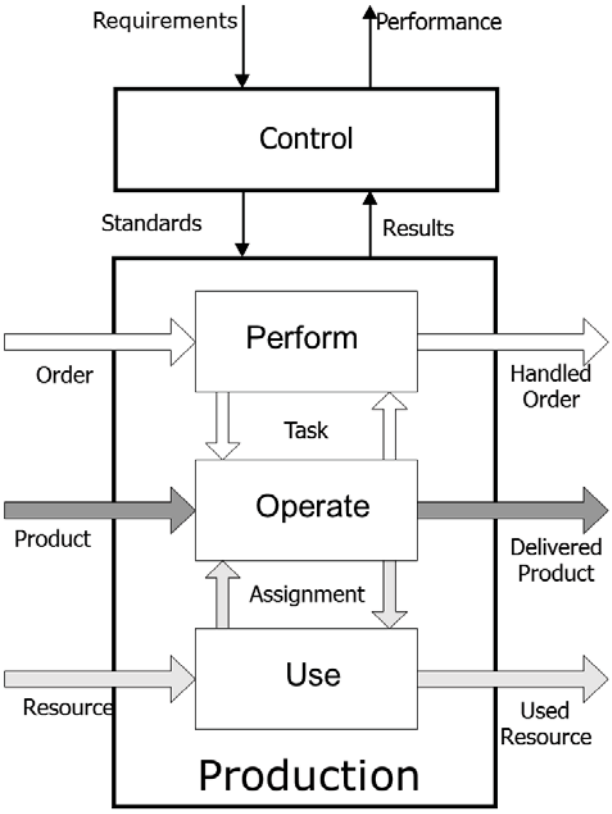


Fig. 2. Proper model

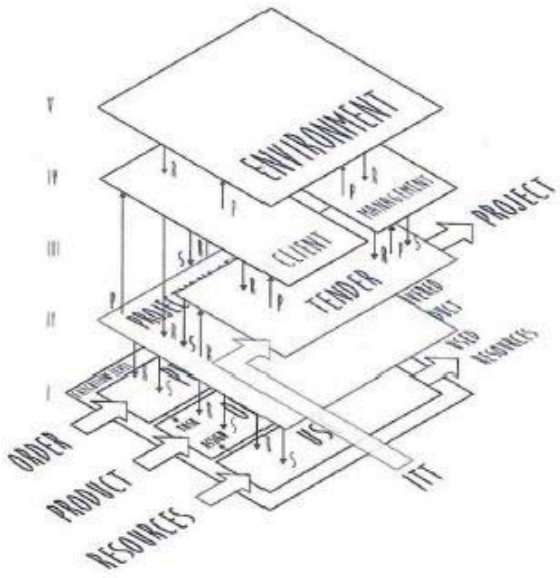


Fig. 3 PROPER-model for the execution of complex projects

At level III an extra perform (user-oriented) function has been incorporated within the control cascade that represents

the tender function. The function embodies the process in which a tender gets transformed into a project, with a certain conversion rate. Subject to the requirements received from management and client, the project specific Δ gets defined based on the basis level S of the executing yard. Performance on the tender process is referred back to client and management to tune the transformation to the (changing) requirements.

The second level involves project control in which the Δ and basis S are set received from the tender function respectively management function. They are translated into project specific standards (Δ +S) for the execution level below. The focus of this function is on the long term function controls, defining project specific standards and evaluating results to act upon possible structural deviations. Long term refers here to the part of the project that overlaps with the tender function.
At the lowest level, Level I, the project execution, three aspect flows are apparent: order, product and resource flow. These flows together transform the input into the output according the project specific standards. In the order flow the order gets translated into clear and executable tasks for the product flow in which the physical transformation/execution takes place. From the resource flow the correct resources are assigned to the product flow.

Framework

In an iterative approach between practice and theory, a framework is developed based on the PROPER model of Fig. 3. The resulting framework is shown in Fig. 4. below. Relations, interdependencies and functions within processes of complex projects become clear and support a controlled steady-state execution.
For the steady-state execution of complex projects both the downward convergence from requirements to standards and finally tasks and assignments have to be improved. The same holds for the upward convergence from execution to results and finally performance. Also the relations and interdependencies have to be defined clearly.
Improving the downward convergence is realized by an appropriate coordination function. The four factors enabling a control function as discussed earlier are incorporated in the framework and provide the necessary project control.

Objectives are formulated within the initiate functions at the different levels. In the perform function at execution level the scope is translated into clear and executable tasks, defining expected outcomes subject to the (Δ + S) standards. By defining and indicating the differences between normal and complex projects with the basis and extra set of standards the framework enables a capable project organization.
The new system allows adjusting its behavior in a pro-active approach; the developed model is empty and conceptual and therefore allows to be adjusted to a variety of projects while keeping the framework similar.
Interdependencies and relationships are defined in the framework and are constant for a variety of projects.

The upward convergence is realized by introducing a learning function with two elements; one being the implementation of a function control in the tender function; the second being the incorporation of an upward results flow by introducing separate evaluate and initiate functions at the different levels within the framework.
The function control within the tender function is the central body of knowledge in the execution of complex projects. By the evaluation of executed projects and the initiation of new possibilities, the Δ is defined and improved in the tender group.

The basis level (S) is provided by the yards, and defined by means of the results of both the execution of complex and normal projects and improved by yard specific developments. To obtain these results, evaluation and compare functions have been incorporated in the model.

All these functions are shown in the model of fig. 4.

These functions serve two purposes. Firstly they should offer a correct convergence of the results to other functions in the upward direction, for up-to-date control and learning; secondly they should enable an efficient transformation within the same function by correctly defining possible deviations on which appropriate actions can be taken by the corresponding initiate function.

The Δ is defined by an analysis of the complex projects, both in expected future and in past experience, and by the basis levels(S) at the different yards. The analysis of executed projects is based on results received from project management and important for the correct definition of the deviation from internal standards. The deviation should be used to improve the standards for the transformation from ITT to Project resulting in a better definition of the Δ set and an improved hand-over.

The learning function and the structured process together will enable continuous improvement of the execution of complex projects. They will finally realize a steady-state execution of a continuous flow of complex projects. The experienced complexity will decrease, according to De Leeuw [2000]. Complexity is a multidimensional concept that concerns interdependency, uncertainty, controllability and heterogeneity; all four will improve by the introduction of the model of Fig. 4.

Results

The framework enables a steady-state execution of a flow of complex projects at different yards. It introduces a singular methodology and centralizes the knowledge by defining the Δ set at the tender group and the basis (S) at the yards. Together they realize a learning function that is verified to prevent up to 14% of the project's cost value. The integrated project control is expected to realize a significant reduction in discrepancy between ambition and realized income.

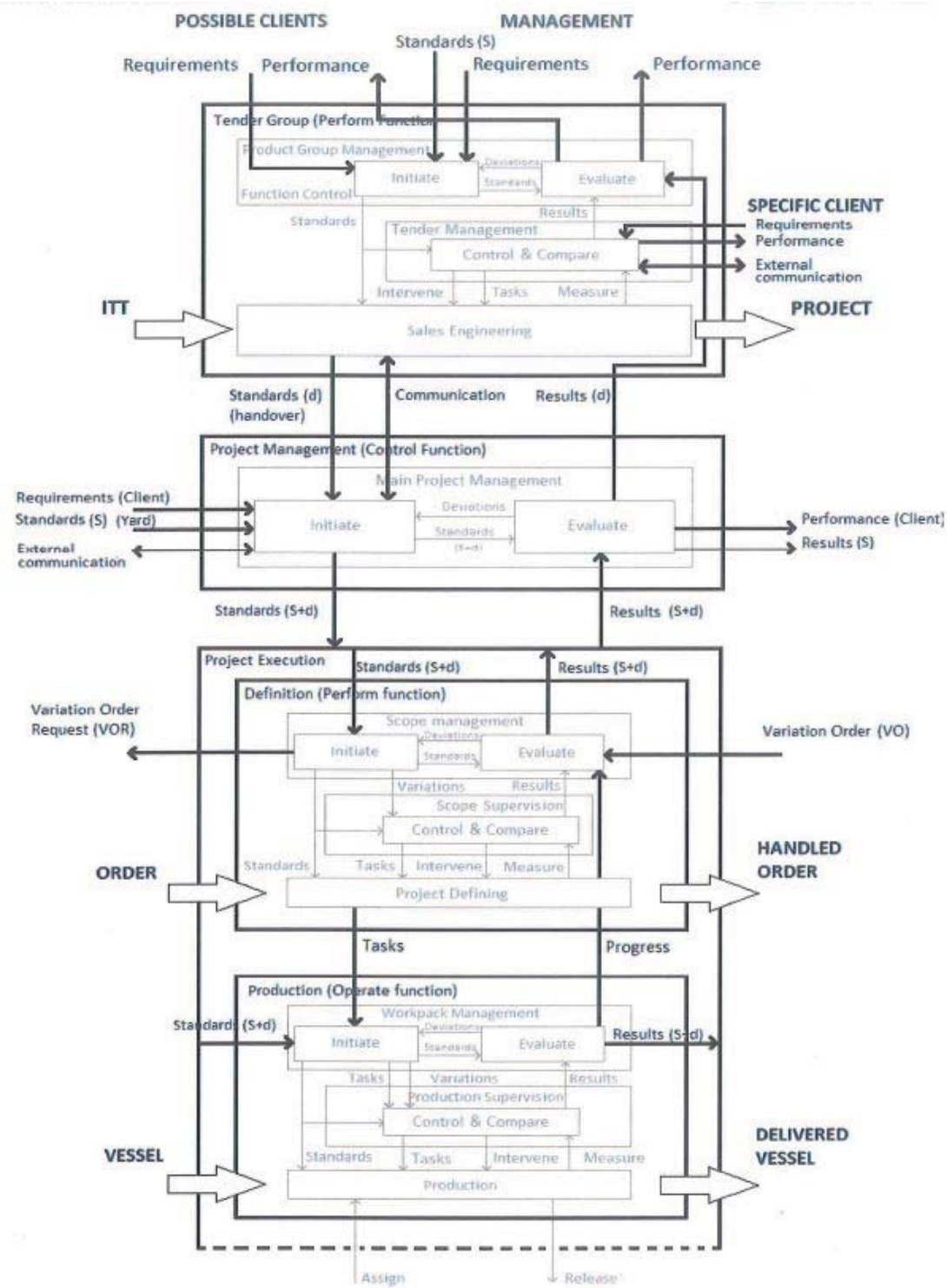


Fig. 4. Functional framework for the improved execution of complex projects

The framework enables the definition of an appropriate project organization for the execution of complex projects.

The framework is functional,, and provides a theoretical basis for the implementation of a singular methodology for conversion yards. Implementing this methodology is currently being done at a yard, and contributes significantly to knowledge gathering and registering experience.

References

[1] Veeke, H.P.M., Lodewijks, G., Ottjes, J.A., "Conceptual Design of Industrial Systems - An approach to support collaboration", 2006, Research in Engineering Design, Vol. 17, Issue 2, pp 85-101
[2] Veeke, H.P.M., Ottjes, J. A., Lodewijks, G., "The Delft Systems Approach", 2008, Springer, ISBN 978-1-84800-176-3
[3] Tushman, M.L., Nadler, D.A., "Information processing as an integrating concept in organizational design", 1978, Ac. Of Management Review, y3, no. 3, pp. 613-624
[4] Nelson, R.R., Winter, S.G>, "An evolutionary theory of economic change", 1982, Harvard University Press, Cambridge
[5] Levering, R., Ligthart, R., Noorderhaven, N., Oerlemans, L., "Continuity and change in inter-organizational project practices: The Dutch Shipbuilding industry", 2013, International Journal of project management, y31, pp. 735-747
[6] Senturk, O.U., "The interaction between the ship repair, ship conversion and shipbuilding industries", 2008, Council working party on shipbuilding
[7] Hall, R.H., "Organizations: Structures, Processes and Outcomes", 1979, New Yersey: Prentice Hall
[8] Baccarini, D., "The concept of project complexity – a review", 1996, International Journal of Project Management, Vil. 14, nr. 4, pp. 201-204
[9] Duncan, W.R., "A guide to the project management body of knowledge", 2013, Boston, Project Management Institute
[10] Kerzner, H., "Project Management. A Systems Approach to planning, scheduling and controlling", 1995, US, Thomson Publishing Inc.
[11] Veld, J. in 't, "Analyse van organisatieproblemen: Een toepassing van denken in systemen en processen", 2002, Houten, The Netherlands, Wolters Noordhoff
[12] Leeuw, A .C.J. de, "Bedrijfskundig management", 2000, The Netherlands, van Gorcum

Acknowledgement
We would like to thank Damen Ship Repair and Conversion for the opportunity to do this research and providing case studies by which valuable data could be gathered and derived.

An augmented pragmatics by explanation-aware and recommendation-aware in the context of decision support

Abdeldjalil Khelassi

Computer science Department, Sciences Faculty, University of Tlemcen, Algeria

Abstract - *The decision support has a strong relationship with pragmatics specifically when the uncertainty is important. In addition, explanation-aware computing is an important component in the process of interaction between Decision Support Systems DSS and users. It represents the basic component of pragmatics for the complex reasoning systems. The recommendation of useful guidelines and research articles is also an important task, which can influence the pragmatics. The intelligent interaction between complex systems and users requires awareness in the context of mass uses applications. The requested awareness can be ensured by including several factors from the context of interaction. In this paper, we present a newest approach for meaningful interaction between a strong Case-Based decision support system and categorized mass users. The originality of this approaches reside in the augmented pragmatics proposed by considering tow aware-component (explainer and recommender).*

Keywords: Decision Support System DSS, Case-based Reasoning, Explanation aware computing, recommendation system.

1 Introduction

The pragmatics is a common concept in several domains the concerned meaning in this work is the Human Machine Interaction and artificial intelligence. In linguistics, pragmatics is the study of the rules and relationships between a language and the interpreters of that language. It is concerned when an investigation explicit reference is made to the speaker, or, to put it in more general terms, to a user of a language [18].

Explanation-aware computing proposes a pragmatic interaction between intelligent systems and users. Artificial Intelligence introduces explanations as a part of reasoning as in the diagnostic applications where the result can be obtained from secondary hypothesis. It considers also explanations in the interaction process to increase the transparency of the obtained results, this side of explanations is due to a request of expert users or special kind of users how need the reasoning process more clear and understandable[19].

For more awareness when using decision support by IK-DCBRC, which is case-based reasoning system for medical Decision support [13-15], we have integrated two components: the explainer and the recommender. We have also defined different levels of abstraction for a meaningful interaction. We have integrated also another component for a transparent deduction of the abstraction levels, according to the user's goals and their category via their social information. In this paper, we present an original proposition for medical application. We explain also the components of this contribution and their dependency.

2 IK-DCBRC for medical decision support

Case-based reasoning is an intelligent approach successfully applied in many medical applications cited in [25] and other domains as information retrieval, Recommender Systems and Decision Support Systems.

IK-DCBRC is a distributed case-based reasoning system for medical decision support [13, 14, and 15]. It contains a set of cognitive agents and an amalgamated knowledge base [16]. We have applied this DSS for cardiac arrhythmias recognize from physiological signals in [14 and 16] the successfully obtained results was published in [35], which is: between 85,5% and 100%, another application was for breast cancer prognosis from cytological image [13 and 15, 36].

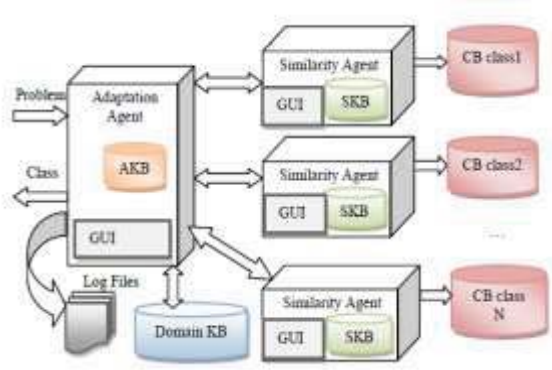


Fig. 1. IK-DCBRC: medical decision support system.

The promising results quality in [13, 14 and 16] encourage us to augment the sophistication of the system. As a first step, we have introduced the explanation-aware as separated agent for explaining the breast cancer diagnosis [15, 36]. This last

option improves significantly the deal with uncertainty problems.

The recommendation systems is a successful application in medical domain, several citations for this success is in [21, 22 and 23]. For this we augment, IK-DCBRC with a recommendation component for more pragmatics.

3 Pragmatics and explanation-aware

As cited before the pragmatics is the study of the rules and relationships between a language and the interpreters of that language. It is concerned when an investigation explicit reference is made to the speaker, or, to put it in more general terms, to a user of a language [18].

In fact, the term explanation has been widely investigated in different disciplines such as cognitive science, artificial intelligence, linguistics, philosophy of science, and teaching. All these disciplines consider certain aspects of the term and make clear that there is not only one such concept but also a variety of concepts [26]. In artificial reasoning systems the pragmatic is ensured by some accessory components as the explainer. Reasoners cited in [26] implement four types of explanation: 1-Reasoning Trace, 2-Justification, 3-Strategic, 4-Terminological. Explanation-aware computing proposes a pragmatic in interaction between intelligent systems and users, which effectively affect the following goals:

Table 1. why explanation is introduced? [27,28]

1. Transparency	How the system reached the answer
2. Justification	Why the answer is a good answer
3. Relevance	Why a question asked is relevant),
4. Conceptualization	Clarify the meaning of concepts
5. Learning	Teach the user about the domain

4 Social networks

As presented in table2, social networking is one of the most popular Web 2.0 applications. The social networking consists that users create a profile and connect with other users as friends or colleagues on numerous different types of free and for-fee sites. Connected people are known as ''friends'' or ''contacts.'' Some sites are primarily professional (e.g., LinkedIn, Researchgate), while other sites are primarily social (e.g., Twitter, Facebook, and MySpace). The social networks and blogs take the biggest portion of the total internet time home and work as explained in table 2. It takes also a biggest portion of internet users on the world for example the Facebook users is, in 2012, 1 billion users on the words and 1,65 billion in 2016. Also this kind of application attracts many social, medical, and didactic investigations as the influence in the education of students and the social learning …etc.

Increasingly, all of these sites include a blend of content. These sites allow for rapid, widespread propagation of information, which provides an opportunity for marketing and social ideas and services, but also has many privacy pitfalls [17, 29].

Table 2. Top 10 online categories by share of total internet time home and work (May 2011)[Nielsen]

Online application	Portion of total internet time home and work
Social networks and blogs	22,59
Online games	9,8
Email	7,6
Portals	4,5
Videos\movies	4,4
Search	4.0
Instant messaging	3,3
Software manufacturers	3,2
Classified\ auctions	2,9
Current events and global news	2,6
Others	35,19

5 Social networks for Recommendation-aware systems

Recommender systems RS principal role is the ability to calculate potential interesting items for users based on their interests. RS are very prevalent for E-Commerce (e.g. Amazon, Netflix), as well as for medical domain [21,22 and 23], and other research area [30, 31, 32, 33]. Three types of recommendation systems are distinguished namely:

- Contents based RS [33] in which the score of recommendation is computed by traditional information retrieval methods as TF-IDF.
- Collaborative based RS [34] in which the application is participative, and the score of recommendations is computed by the users rating.
- Hybrid RS [31] this method combine the previews methods for increasing the awareness's of the recommendation.

In social networks, where each user can be categorized and identified, it is an open challenge to select proper recommenders for predicting the trustworthiness of online information. As a heuristic used in real life, people who are close and influential to us can usually make more proper and acceptable recommendations [24], also professional categorized persons are trusted in their specialty. Based on these observations, we present the idea of this contribution toward, by the recommendation-aware.

We have realized several experiments for user's category identification from profiles patterns in [37, and 38] with and without missing data imputation. The figure2 shows three experiments with three strategies (Expectation-maximization algorithm, Hierarchical Clustering, and K-means clustering) (without missing data, Imputation with KNN, and imputation with mean). The population contains social information about 275 users collected and published by Eugene Dubossarsky and Mark Norrie 2004.

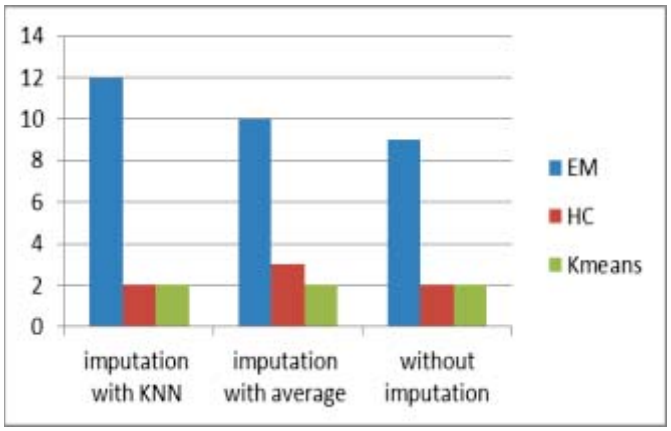


Fig. 2. Categorization of users by social identities and clustering methods (results represented by number of clusters).

The analyze of results was difficult but improve that k-means method keep the same number of clusters, the HC is close to the k-means but no relevant results with the EM method.

Other method was presented in [37] the most transparent one is the rule-based categorization, in which we use a set of rules about users' features to detect the identity.

Another conceptual research work presented in [38]. The work presents the construction of federated identity via an algorithm of data integration from several social networks.

6 Contribution for an aware pragmatism

The context of our application is not just to provide more pragmatic for the complex applications, but also to resolve the problem of information inconsistency by ensuring an adaptive and participative interaction between the users, doctors and the intelligent system for decision support, by explanations and recommendations. An online medical decision support system by IK-DCBRC is developed, for mass uses. In this paper, we introduce two components for an augmented pragmatism and adaptability. For trust achievements, more constraints are added as categorization of users via social networks, and weighted recommendation of documents according to the recommender category.

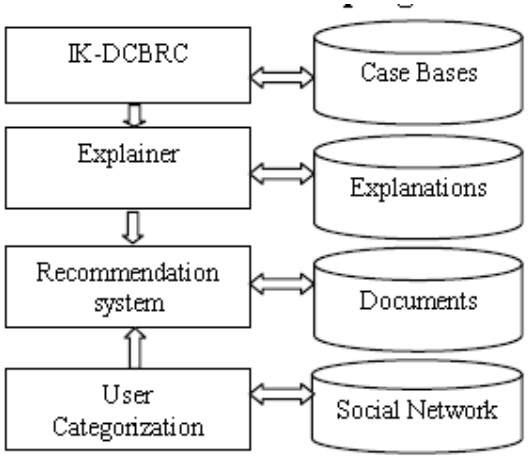


Fig. 3. The application components

An original medical application is developed for medical decision support and explanation (see figure 3 and figure 4). The first task by this application is the categorization of users via their information extracted from profiles in social networks. The user category is very important for the explainer and the RS. The user interacts with the IK-DCBRC decision support system by introducing a query described by a pattern (see figure 4), which describes the features of this query.

Following that, the system generates the disease by inferring from the amalgamated knowledge base and by remembering in the stored cases, more details are in [13, 14 and 16]. A separated component called explainer generates the appropriate explanation to the appropriate user according to the assigned level of abstraction, more details about the explainer are in [15].

The RS collect a query from explanations and the user category for describing the context of the query. This increases significantly the awareness of the recommendation, but another process is also solicited by analyzing the contents of the indexed documents based on tf-idf approach see [39] for more details.

The RS also consider the feedback of users in all categories, but each category has a weight for their recommendations. Another new constraint added to the feedback step, the selection of the category for the recommended item (see figure 4).

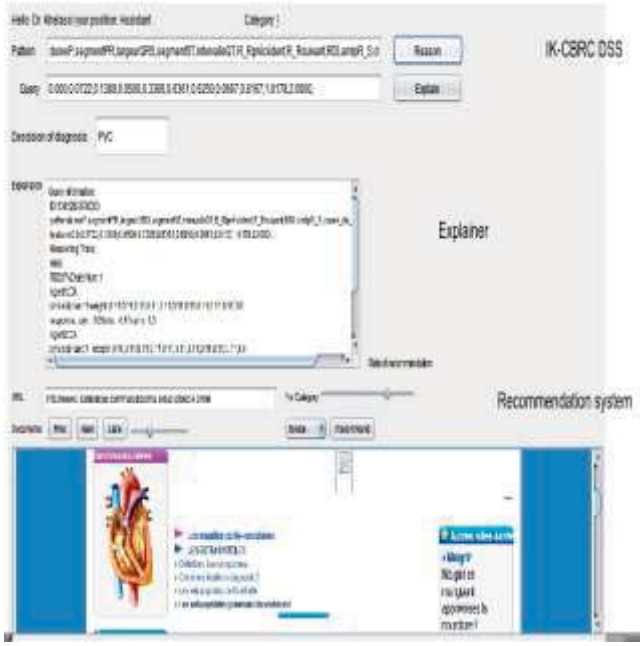


Fig. 4. Document recommendations after medical diagnosis and explanation

The developed system contains an index of some e-documents for the novice users and scientific articles for the professionals (100 documents from Pubmed). The index contains the link of the document associated with the key words, the kind of readers and the rate of recommendation. More details about the recommender component are in [39].

For resolving the problem of online information, which can be inaccurate, incomplete, controversial, misleading, and alarming for individuals with health questions [8, 20]. The users' feedbacks and recommendations control the quality of information by considering the category of the recommenders. For example, the recommendation of document by a doctor for a novice user is more rated than the recommended by the novice user for novice user. Some rules are defined for storing and defining the most relevant rate of recommendation.

For measuring the effectiveness of explanation-aware component, we have collected the appreciation of users by using the same dataset used in [14 and 16]. The figure5 presents the results of users appreciation.

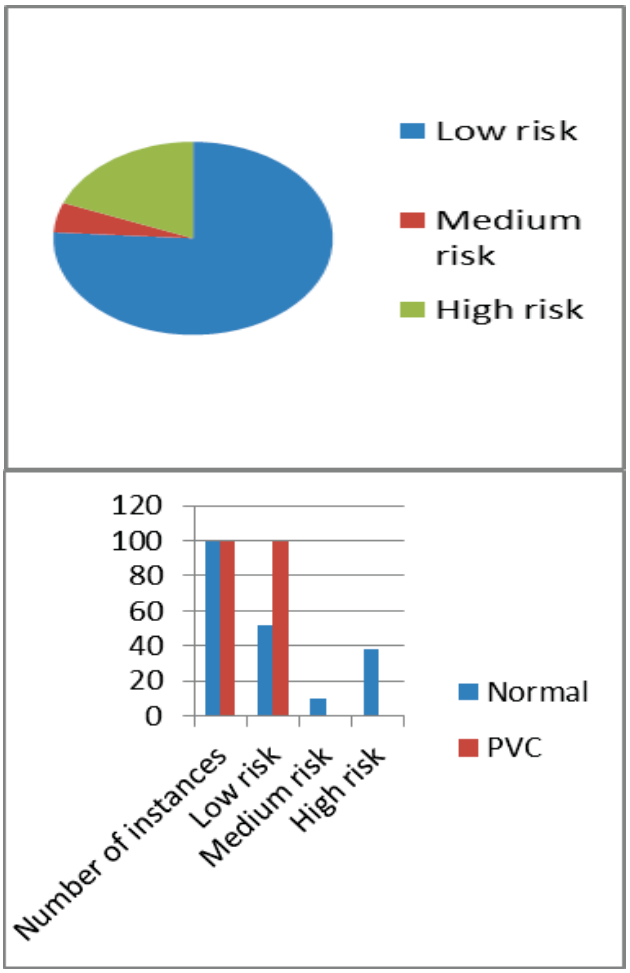


Fig. 5. Appreciation of users after explanation visualization.

7 Conclusion

The goal of this work is to propose an original online medical decision support system enriched by explanation-aware computing and augmented by RS. This contribution ensures the needed transparency and increases the information quality by the recommendation-aware proposition.

This work achieves the aim of the proposition but there are some important limits as the credibility of the social information, the biggest quantity of documents. Our perspectives focus on ensuring the smartest interaction between the decision support system and the users.

8 REFERENCES

[1] INTERNET USAGE STATISTICS juin 2012 http://www.internetworldstats.com/stats.htm

[2] Fox, S. (2009). The social life of health information. www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx

[3] Clark, N. M. (2003). Management of chronic disease by patients. Annual Review of Public Health, 24, 289-313

[4] Barker, K. K. (2008). Electronic support groups, patientconsumers, and medicalization: The case of contested illness. Journal of Health and Social Behavior 49(March):20-36

[5] Eysenbach, G. (2003). The impact of the internet on cancer outcomes. CA: American Cancer Journal for Clinicians,54:356-371

[6] Farnham, S., Cheng, L., Stone, L., Zaner-Godsey, M., Hibbeln, C., Syrjala, K., et al. (2002). Hutchworld: Clinical study of computer-mediated social support for cancer patients and their caregivers. CHI'02, pp. 375-382

[7] Cummings, J., Sproull, L. S., & Kiesler, S. (2002). Beyond hearing: Where real-world and online support meet. Group Dynamics: Theory, Research and Practice, 6(1):78-88

[8] Glynn et al. BMC Cancer 2011, The effect of breast cancer awareness month on internet search activity - a comparison with awareness campaigns for lung and prostate cancer 11:442 http://www.biomedcentral.com/1471-2407/11/442

[9] Burke, M., Kraut, R., & Williams, D. (2010). Social use of computer-mediated communication by adults on the autism spectrum. CSCW'10, 425-434

[10] Kummervold, P. E., Gammon, D., Bergvik, S., Johnsen, J., Hasvold, T., & Rosenving, J. (2002). Social support in a wired world: Use of online mental health forums in Norway. Nordic Journal of Psychiatry , 56:59-65

[11] Jennifer Mankoff, al Competing Online Viewpoints and Models of Chronic Illness, CHI 2011 • Session: Health 3: Online Communities & Social Interaction May 7–12, 2011 • Vancouver, BC, Canada.

[12] Sue Black, Joanne Jacobs, Using Web 2.0 to Improve Software Quality, Web2SE '10 Proceedings of the 1st Workshop on Web 2.0 for Software Engineering, ACM New York, NY, USA ©2010

[13] A.khelassi: ''Data mining application with case-based reasoning classifier for breast cancer decision support''. MASAUM Inter Conf on Info Tech 2012 MI-CIT'12; 07/2012

[14] Abdeldjalil KHELASSI, Mohamed Amin Chick. Fuzzy knowledge-intensive case-based classification for the detection of abnormal cardiac beats. Electronic Physician, 2012;4(3):565-571.

[15] KHELASSI, Abdeldjalil. Explanation-aware computing of the prognosis for breast cancer supported by IK-DCBRC: Technical innovation. Electronic Physician, 2014, vol. 6, no 4, p. 947.

[16] Khelassi Abdeldjalil, Mohammed Amine Chikh. Cognitive Amalgam with a Distributed Fuzzy Case-based Reasoning system for an accurate cardiac arrhythmias diagnosis. International Journal of Information and Communication Technology. 2015. 7(4/5):348-365.

[17] Khelassi Abdeldjalil. RAMHeR: Reuse And Mining Health2.0 Resources. Electronic Physician. 2015. 7(1):969-970.

[18] Mark Dietrich Tschaepe. Pragmatics and Pragmatic Considerations in Explanation. Contemporary Pragmatism.Vol. 6, No. 2 (December 2009), 25–44

[19] Agnar Aamodt. A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning. PhD thesis, Norwegian Institute of Technology, Department of Computer Science, Trondheim, May 1991.

[20] Matthew S. Katz, al "The 'CaP Calculator': an online decision support tool for clinically localized prostate cancer BJU International" Volume 105, Issue 10, pages 1417–1422, May 2010, DOI: 10.1111/j.1464-410X.2010.09290.x

[21] CHEN, Rung-Ching, HUANG, Yun-Hou, BAU, Cho-Tsan, et al. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications, 2012, vol. 39, no 4, p. 3995-4006.

[22] BACHUS, Sonja, ABREU, Michael, BACHUS, Jeff, et al. Healthcare provider recommendation system. U.S. Patent Application 11/181,158, 13 juill. 2005.

[23] HELLWIG, Robert et WEINERT, Stefan. Insulin bolus recommendation system. U.S. Patent No 7,291,107, 6 nov. 2007.

[24] JIANG, Wenjun, WU, Jie, et WANG, Guojun. RATE: Recommendation-aware Trust Evaluation in Online Social Networks. In : Network Computing and Applications (NCA), 2013 12th IEEE International Symposium on. IEEE, 2013. p. 149-152.

[25] I. Bichindaritz, C Marling, Case-based reasoning in the health sciences: What's next? Artificial Intelligence in Medicine 36 (2), 127-135.

[26] Thomas R. Roth-Berghofer, Stefan Schulz, and David B. Leake, editors, explanation-Aware Computing – Papers from the 2007 AAAI Workshop, number WS-07-06 in Technical Report, pages 20–27, Vancouver, BC, 2007. AAAI Press.

[27] Thomas R. Roth-Berghofer and Jörg Cassens "Mapping Goals and Kinds of Explanations to the Knowledge Containers of Case-Based Reasoning Systems"Héctor Muñoz-Avila and Francesco Ricci, editors, Case-based Reasoning Research and Development – ICCBR 2005, volume 3630 of LNAI, pages 451– 464, Chicago, 2005. Springer.

[28] Jörg Cassens and Anders Kofod-Petersen, "Explanations and Case-Based Reasoning in Ambient Intelligent Systems" David C. Wilson and Deepak Khemani, editors, ICCBR-07 Workshop Proceedings, pages 167–176, Belfast, Northern Ireland, 2007.

[29] Julia C. Phillippi, al Web 2.0: Easy Tools for Busy Clinicians Volume 55, No. 5, September/October 2010 1526-9523.

[30] Burke, R. 2000. Knowledge-based recommender systems. In: Dekker, M. ed. Encyclopedia of Library and Information Systems 69. New York, NY, USA. 180–200.

[31] Burke, R. 2007. Hybrid Web Recommender Systems. In Brusilovsky, P., Kobsa, A., Nejdl, W. eds. The Adaptive Web, LNCS 4321. Springer Berlin/Heidelberg, 377–408.

[32] Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S.2007. The adaptive web. Springer Berlin/Heidelberg. 291-324

[33] Pazzani, M. J. and Billsus, D. 2007. Content-based recommendation systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W. eds. The adaptive web. LNCS 4321. Springer-Verlag, Berlin, Heidelberg. 325–341.

[34] de Campos, L. M., Fernández-Luna, J. M., & Huete, J. F. (2008). A collaborative recommender system based on probabilistic inference from fuzzy observations. Fuzzy Sets and Systems, 159(12), 1554-1576.

[35] KHELASSI, Abdeldjalil. The impact of uncertainty measures in the cardiac arrhythmia detection supported by IK-DCBRC. In : Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015. p. 144.

[36] KHELASSI, Abdeldjalil. The Effect of Explanation-aware Computing in Breast Cancer Detection In Conference: BIT's 8th Annual World Cancer Congress, At Busan, Republic of Koria Dec 2015.

[37] H Naim (2014) Estimation des données manquantes et catégorisation des identités sociaux, Master thesis UABB Tlemcen

[38] DIB, Sidi Mohammed Fazil et LARABI, Nor El Islam. Construction d'une identité sociale fédérée représenté en XML par une approche d'intégration de données. 2014. Master thesis UABB Tlemcen.

[39] TCHENAR, Med Ilyas et RAHALI, Youssouf. Une application médicale de recommandation contextuelle des documents. 2014. Master thesis UABB Tlemcen.

Intelligent Power Oscillation Damping Control with Dynamic Knowledge Inference

R. K. Pandey, *Senior Member IEEE*
Department of Electrical Engineering
IIT (BHU), Varanasi
India
rpsneh@yahoo.co.in

Deepak Kumar Gupta, *Student Member IEEE*
Department of Electrical Engineering
IIT (BHU), Varanasi
India
dpkgpt214@gmail.com

***Abstract*— The paper presents an intelligent power oscillation damping control with dynamic knowledge inference concept. The power oscillation damping in an interconnected large power network requires controllers suitable deployed and regulated with changing system conditions. The combined intelligent control strategy has been proposed for real and reactive power regulation. The Gravitational Search Algorithm (GSA) optimization technique has been used to get the optimal parameters of respective controllers and this has been used in developing knowledge domain with dynamic inference. A Sample six area system has been considered to demonstrate the controlled states with operational shift.**

Keywords*— Gravitational Search Algorithm (GSA), Integral time multiplied by absolute error (ITAE), Knowledge Inference Mechanism (KIM), Power System Stabilizer (PSS), Static Synchronous Compensator (STATCOM).

I. INTRODUCTION

Damping in the power oscillation is desirable as it not only reduces variation in system states and improves the power quality but also enhances stability limits. As power demand grows rapidly with increase in transmission line length and generators with limited availability of resources, the existing power systems are frequently loaded beyond nominal range. It is known that a supplementary controller (PSS) is normally used to provide damping and improve dynamic performance. However, PSSs may fail to stabilize for larger perturbation, under such situation, FACTS controllers have shown promising results to improve oscillation damping. STATCOM is a regulating device based on a power electronics voltage source converter and can act as either reactive power injection or absorption in the network [1-4]. Many intelligent optimization techniques have been used to design these controllers. Abdel-Magid YL, Abido MA et al used genetic algorithm and tabu search techniques to design power system stabilizers for multi-machine power systems [5-6]. Eslami M et al also reported use of GA and PSO techniques for power oscillation damping [7-8]. Small signal stability model for FACTS devices (UPFC, STATCOM and SSSC etc.) have been reported in [9-10]. PSO technique has been used to tune SVC and PSS [11]. In [12-15], TCSC and UPFC have been designed with intelligent optimization techniques (GA, PSO and Quantum Particle Swarm Optimization) to damp out inter-area oscillation in the system. R. K. Pandey et al. reported UPFC control parameter identification for effective and precise power oscillation damping and the design of multi-stage LQR control strategy, which results in optimal tracking. The design of intelligent weight matrix has been introduced for Q matrix based on the state predominant approach [16-17].

This paper presents further work reported as in [18-19] which gives the concept of controller shifting/sharing as the system operating condition changes with time. Controller shifting concept basically operates within the controller by redesigning the control parameters of respective controller with the change in system operating conditions. Controller sharing concept is realized when local controllers (PSS) reach their maximum capacity and unable to stabilize the system at certain operating conditions. In that case any controller connected nearby in the system (STATCOM) will act as a supplementary controller in addition to the existing one and stabilize entire system quickly. To develop knowledge linked inference mechanism for all controllers connected in the network, a GSA optimization technique has been used. This provides a new control structure and helps in quick regulation with precise damping. However, while realizing such structure in field, this may require additional intelligent soft controllers which may have multi-controller parameters realization in given knowledge domain framework. A sample six area system has been considered with STATCOM connected between area 3 and area 4 and all the generators with PSS of different ratings. Multi-stage LQR control concept is used to design the input control parameters of STATCOM [17].

II. STATE SPACE MODELING OF PSS AND STATCOM

A. Approach for PSS model inclusion

Figure 1 shows the multi machine representation with n number of generators connected in each area. Vt1, Vt2 are the terminal voltage, I1, I2 are the armature current and Y1, Y2 represents loading in respective area. Z represents transmission line.

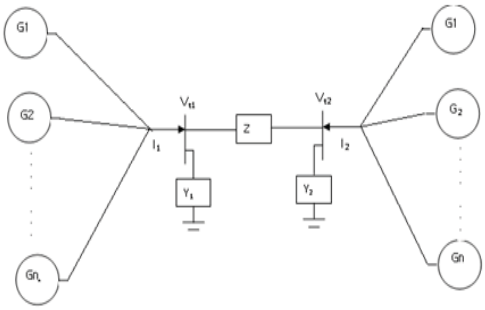


Figure. 1 Multi Machine Representation

Armature current and terminal voltage can be represented by:

$$i_1 = i_{d1} + ji_{q1}\ ,\ v_{t1} = v_{d1} + jv_{q1} \quad (1)$$

$$v_{di1} + jv_{qi1} - j * x_{ti1} * (i_{d1} + ji_{q1}) = v_{d1} + jv_{q1} \quad (2)$$

Following constant and parameters are introduced for convenience.

$$C11 = 1 + R * G1 - X * B1,\ C21 = X * G1 + R * B1 \quad (3)$$

$$1 + ZY1 = C11 + jC21, \quad (4)$$

$$R'_{i11} = R - C21 * x''_{di1}, R'_{i21} = R - C21 * x'_{qi1} \quad (5)$$

$$X'_{i11} = X + C11 * x'_{qi1}, X'_{i21} = X + C11 * x''_{di1} \quad (6)$$

$$Z^2_{ei1} = R'_{i11} * R'_{i21} + X'_{i11} * X'_{i21} \quad (7)$$

$$Y_{di1} = (C11 * X'_{i11} - C21 * R'_{i21}) / Z^2_{ei1}, \quad (8)$$

$$Y_{qi1} = (C11 * R'_{i11} + C21 * X'_{i21}) / Z^2_{ei1}, \quad (9)$$

Where Z=R+jX (Trans. Line Reactance) and Y=G+jB (Load). From Fig. 1 we will have:

$$i_1 = Y1 * v_{t1} + Z^{-1}(v_{t1} - v_{t2}) \quad (10)$$

$Z * i_1 = (1 + Z \,\&\, Y1) * v_{t1} - v_{t2}$, this can be written as:

$$\begin{bmatrix} R & -X \\ X & R \end{bmatrix} \begin{bmatrix} i_{d1} \\ i_{q1} \end{bmatrix} = \begin{bmatrix} C11 & -C21 \\ C21 & C11 \end{bmatrix} \begin{bmatrix} v_{d1} \\ v_{q1} \end{bmatrix} - v_{t2} \begin{bmatrix} Sin(\delta_{im} - \delta_{jn}) \\ Cos(\delta_{im} - \delta_{jn}) \end{bmatrix} \quad (11)$$

From equation (2) and equation (4):

$$\begin{bmatrix} i_{d1} \\ i_{q1} \end{bmatrix} = \begin{bmatrix} Y_{di1} \\ Y_{qi1} \end{bmatrix} E'_{qi1} - (v_{t2} / Z'^2_{ei1}) \begin{bmatrix} R'_{i21} & X'_{i11} \\ -X'_{i21} & R'_{i11} \end{bmatrix} \begin{bmatrix} Sin(\delta_{im} - \delta_{jn}) \\ Cos(\delta_{im} - \delta_{jn}) \end{bmatrix} \quad (12)$$

$$\Delta\delta = \delta_{im} - \delta_{jn} \quad (13)$$

$$\begin{bmatrix} \Delta i_{d1} \\ \Delta i_{q1} \end{bmatrix} = \begin{bmatrix} Y_{di1} \\ Y_{qi1} \end{bmatrix} \Delta E_{qi1} + \begin{bmatrix} F_{di1} \\ F_{qi1} \end{bmatrix} \Delta\delta \quad (14)$$

Where:

$$\begin{bmatrix} F_{di1} \\ F_{qi1} \end{bmatrix} = -(v_{t2} / Z^2_{ei1}) \begin{bmatrix} R_{i21} & X_{i11} \\ -X_{i21} & R_{i11} \end{bmatrix} \Delta E_{qi1} + \begin{bmatrix} F_{di1} \\ F_{qi1} \end{bmatrix} \Delta\delta \quad (15)$$

Final states equations for one machine can be written as:

$$Ms\Delta\omega = -\Delta T_e = -(K_1\Delta\delta + K_2\Delta E'_q) \quad (16)$$

$$s\Delta\delta = \omega_b \Delta\omega \quad (17)$$

$$(1 + sT'_{do}K_3)\Delta E'_q = K3(-K_4\Delta\delta + \Delta E_{fd}) \quad (18)$$

$$(1 + sT_A)\Delta E_{fd} = K_A(u_E - \Delta v_t) = K_A(u_E - K_5\Delta\delta - K_6\Delta E'_q) \quad (19)$$

Using above four equations, the state variables vector becomes:

$$x = [\Delta\omega, \Delta\delta, \Delta E'_q, E_{fd}] \quad (20)$$

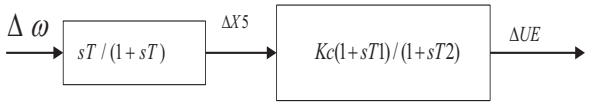


Figure. 2 PSS Model

Figure 2 shows the block diagram for the power system stabilizers where one block is the phase lead-lag compensation and the second one is the reset block which used to activate the supplementary excitation when system oscillation begins.

$$(1 + sT)\Delta X5 = sT * \Delta\omega \quad (21)$$

$$(1 + sT2)\Delta UE = Kc(1 + sT1)\Delta X5 \quad (22)$$

Final states space equation combining PSS is:

$$\dot{\Delta X} = [A]\Delta X + [B]\Delta UE = [Ac]\Delta X \quad (23)$$

Where B is the control matrix, ΔUE the supplementary excitation and [Ac] is the controlled system matrix. Where:

$$\Delta X = \left[\Delta\delta, \Delta\omega, \Delta E'_q, \Delta E_{fd}, \Delta X5, \Delta UE\right]^T \quad (24)$$

B. Approach for STATCOM model inclusion

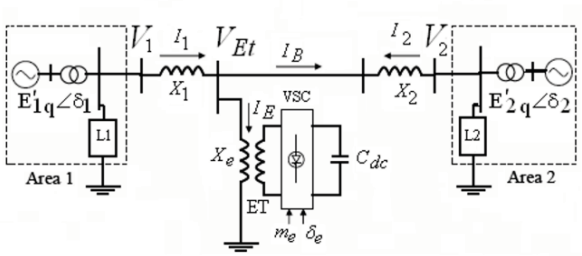


Figure. 3 STATCOM Installed Between Two Area

STATCOM connected between area 1 and 2 is shown in figure 3 using Thevenin's equivalent model. STATCOM is composed of DC-link capacitor, GTO base voltage sources converters, and excitation transformer (ET) connected in shunt. PWM technique has been considered for developing the model of STATCOM. Input control parameters are amplitude modulation ration (Δm_e) and phase angle ($\Delta\delta_e$) for voltage source converter.

The linearized equations of two area power system are [10]:

$$\dot{\Delta\delta} = \omega_0\Delta\omega \quad (25)$$

$$\dot{\Delta\omega} = \frac{-\Delta P_e - D.\Delta\omega}{M} \quad (26)$$

$$\dot{\Delta E'_q} = \frac{-\Delta E'_q - (X_d - X'_d)\Delta I_d + \Delta E_{fd}}{T'_{d0}} \quad (27)$$

$$\dot{\Delta E}_{fd} = \frac{-\Delta E_{fd} + K_a(-\Delta V_t)}{T_a} \quad (28)$$

The complete state space model of two area power system installed with STATCOM can be obtained as following:

$$\dot{\Delta x} = A\Delta x + B\Delta u \quad (29)$$

$$\Delta x = \left[\Delta\delta, \Delta\omega_1, \Delta\omega_2, \Delta E'_{1q}, \Delta E_{1fd}, \Delta E'_{2q}, \Delta E_{2fd}, \Delta V_{dc}\right]^T \quad (30)$$

$$\Delta u = \left[\Delta m_e, \Delta\delta_e\right]^T \quad (31)$$

III. Gravitational Search Algorithm (GSA)

GSA is a newly developed intelligent search technique based on the gravitational law and interaction of masses. In GSA, agents are collection of masses and considered as the object (candidate solutions) with their performance measured by their masses. All objects/agents attract each other by the gravity force resulting in a global movement of all objects towards the objects with heavier masses [20].

Consider a system with N number of objects with d-dimensions, then position of each agent i is defined as:

$$X_i = (X_i^1, ...X_i^d, ...X_i^n) \text{ for i=1,2,…N,} \quad (32)$$

The force acting on mass 'i' due to mass 'j' is:

$$F_{ij}^d(t) = G(t)\frac{M_{pi}(t) * M_{aj}(t)}{R_{ij}(t) + \varepsilon}(x_j^d(t) - x_i^d(t)) \quad (33)$$

R_{ij} is the Euclidian distance between agent i and agent j.

$$R_{ij}(t) = \| X_i(t), X_j(t) \|_2 \quad G(t) = G(G_0, t) \quad (34)$$

Where, $G(t)$ is the gravitational constant, M_{aj} is the active gravitational mass of j agent and M_{pi} is the passive gravitational mass related to agent i. ε is small constant.

Total force acts on a single agent i and acceleration in a dimension d from other agents is:

$$F_i^d(t) = \sum_{j=1, j\neq i}^{N} rand_j F_{ij}^d(t), \; a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \quad (35)$$

where, M_{ii} is the inertial mass of i[th] agent.

Velocity of agent is updated by:

$$v_i^d(t+1) = rand_i * v_i^d(t) + a_i^d(t) \quad (36)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (37)$$

And the masses of each agent are updated by:

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)}, \; M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \quad (38)$$

Here $best(t)$ and $worst(t)$ both are depend upon objective problem (i.e., maximization or minimization).

IV. Knowledge Inference Mechanism (KIM)

Knowledge inference mechanism can be defined as the set of rules for designing the controller's parameters at operational shift in the network condition. With the change in the perturbation in the network, oscillation in the states variables will increase and may violates form the desired limits or may sustain for the longer time which finally will affect the stability of the system. Damping of these oscillations depends upon the setting of controller parameters and if the design of controller is not proper then it will not only degrade the system performance but also aggravate the adverse dynamic conditions. Knowledge domain is the set of tuned controller parameters at different operating condition which is developed and stored in off-line conditions. It includes the range of tuned control parameters in which systems state variables are within their limits. Development of this knowledge domain can be done by any heuristic optimization technique with any objective function, which should be related to the damping in the system. In this paper GSA optimization techniques is considered and ITAE is taken as objective function [18-19].

The objective function which is used to generate Knowledge Domain by the use of gravitational search algorithm is ITAE (minimization problem).

$$J\{e(t)\} = ITAE = \int_0^\infty t^* |e(t)| dt \quad (39)$$

Where e(t) is the error of the state variables from their desired values. GSA is used to tune the control parameters for the PSS and STATCOM for many system operating conditions off-line. All the tuned parameters are stored in their respective knowledge domain.

A. Knowledge Retrieval with Inference Mechanism and Controllers realization

Change in the system operating condition is linked with the percentage change in the system state variables response and as the dynamical operational shift occurs, corresponding change in the behavior of the system is measured. Values of controllers parameters stored in the knowledge domain are linked with the respective operational change in the network. With the detection of percentage change in the behavior of the system, retuning of controller take place with the help of knowledge inference mechanism with some switching delay and stabilize the system as quickly as possible. (Figure. 5)

B. Complete Regulation of the System

Complete system regulation has to be checked after retuning the controller's parameters and modulating the power flow in the network. There might be some operating condition where all the controllers connected in the network will not perform satisfactory, in that case isolation of that particular region from the system completely will serve as a final control action (may be load shedding or SPS operation) and prevent major collapse in the system.

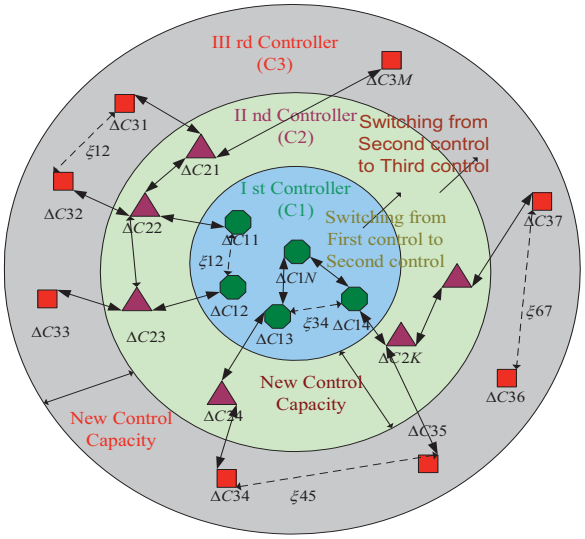


Figure. 4 Control Structure for Knowledge Domain Mapping

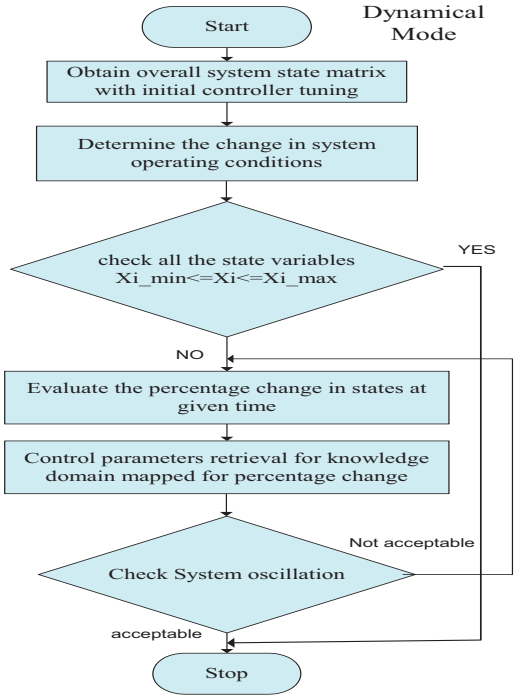


Figure. 5 Flow chart for controller realization in dynamical mode

V. Case Study and Results

Two cases have been considered to demonstrate the effectiveness of controller sharing and shifting concept with GSA driven knowledge inference mechanism. First case represents the retuning of PSS controllers parameters (controller shifting) as the system operating condition changes with time and retuned parameters damp out the oscillation more as compared to the previous tuned values. Second case represents the effectiveness of the STATCOM connected in the network when PSS fails to stabilize the network at certain system operating condition. PSS with STATCOM together modulated the power flow in the network as desired and brings the system stable. Six Area Systems have been considered with STATCOM connected between area 3 and area 4 and PSS to all the areas. Time constants (T1 and T2) of lead-lag compensation block are used as the control parameters for PSS and amplitude modulation ratio (m_e) and phase angle (δ_e) as the control parameters for STATCOM. Table 1 shows the comparison between system responses in terms of overshoot/undershoot and settling time for tuning of controller with GSA driven knowledge domain inference mechanism concept and without this concept. Figure 7 shows the system states variables response with GSA driven KIM based retuned PSS parameters for change in operating condition which enhance the oscillation damping in the system and stabilize the network. Figure 8 shows some operating condition where PSS only not able to stabilize the system, in that case STATCOM acts as supplementary controller in addition to the existing PSS controller and stabilize the system as quickly as possible (shown in Figure. 9). The results demonstrate that over-shoot/ under-shoot and settling time of the system state variables are greatly reduced by applying the proposed concept for tuning of controllers.

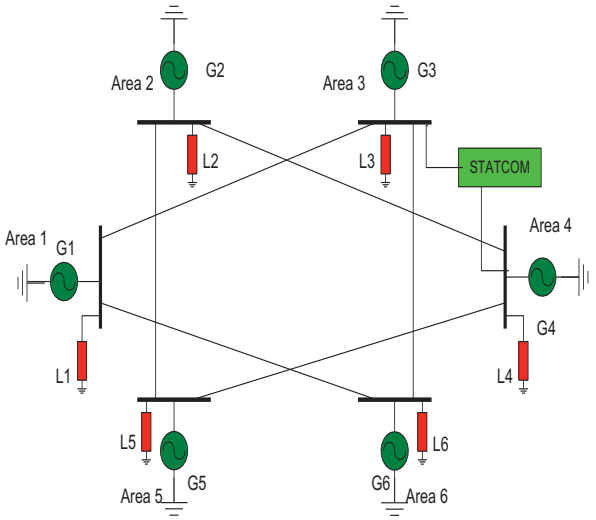


Figure. 6 Six Area System with STATCOM between Area 3 and 4

Load in Area 3 (Y3)=G3+j*B3; where G3=1.6615; B3=1.9029;

Case I- L3(1)=1.6*Y3 at time t1 and L3(2=0.8*Y3 at t2; where t1=0sec and t2=1.5 sec

Case II=L3(3)=1.7*Y3 (**1-** PSS only / **2-** STATCOM supplements PSS)

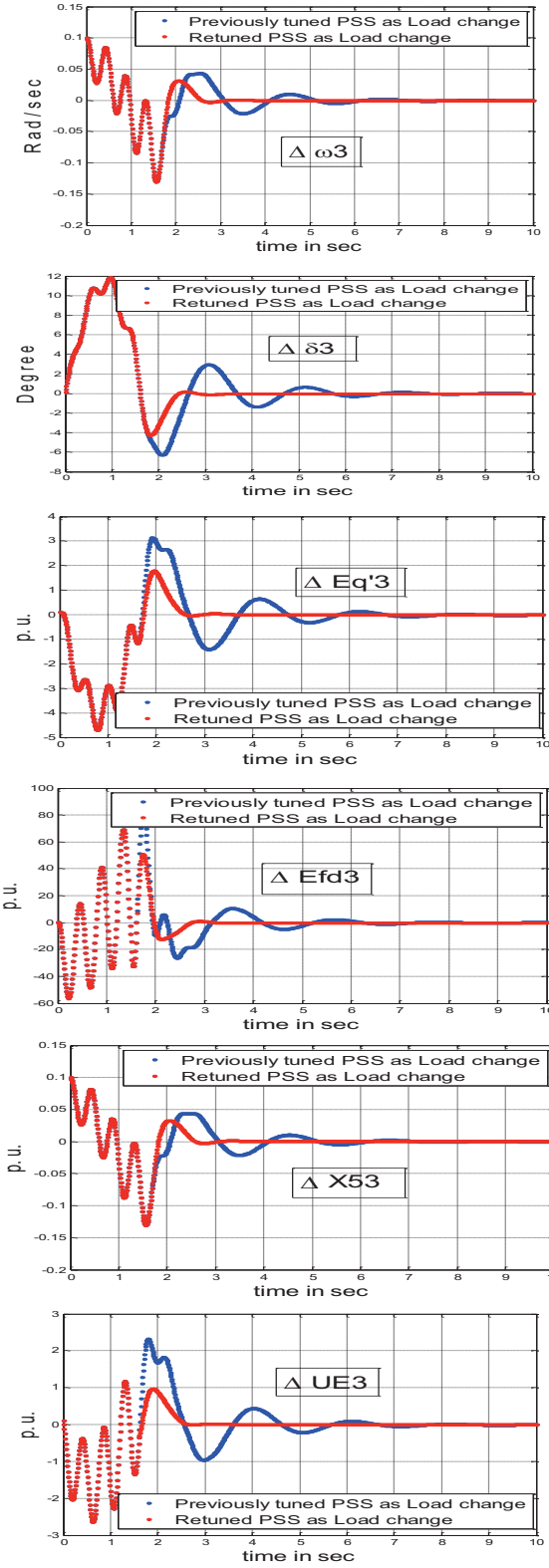


Figure. 7 Perturbation response of system states for Case I for Six Area System (Area 3)

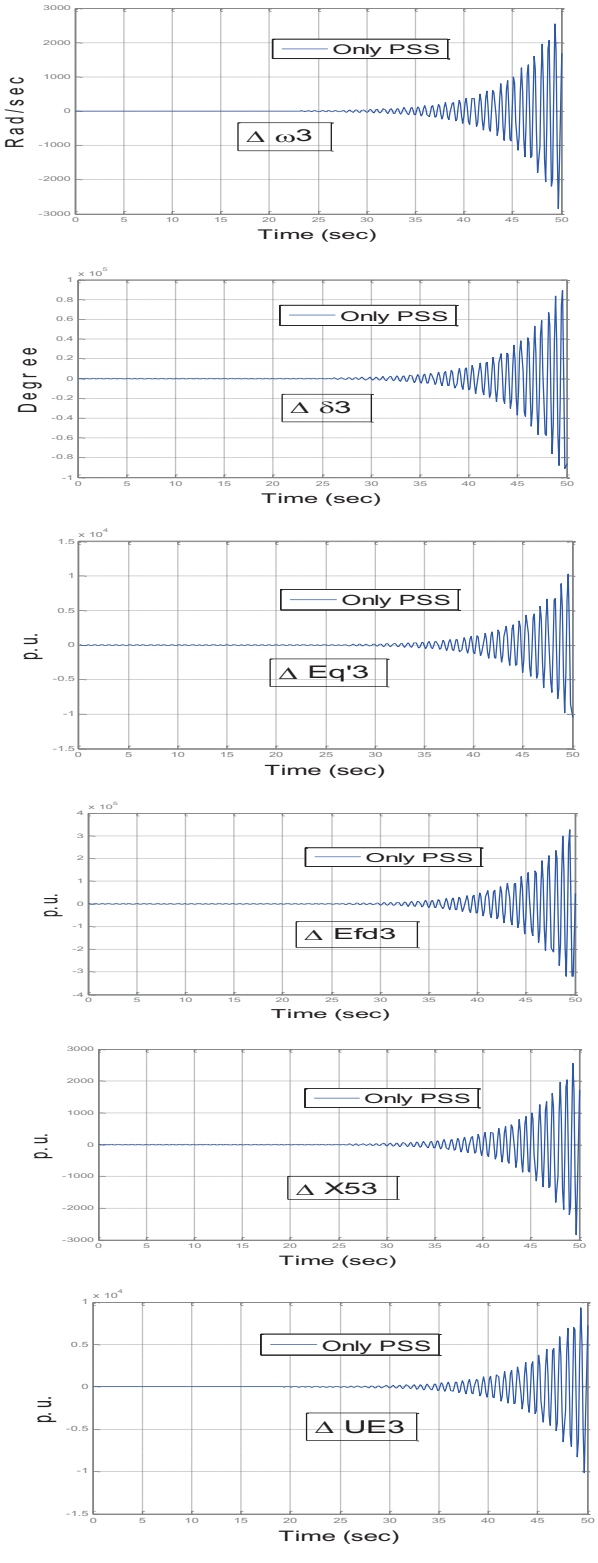


Figure. 8 Perturbation responses of system states for Case II for Six Area System only with PSS (Area 3)

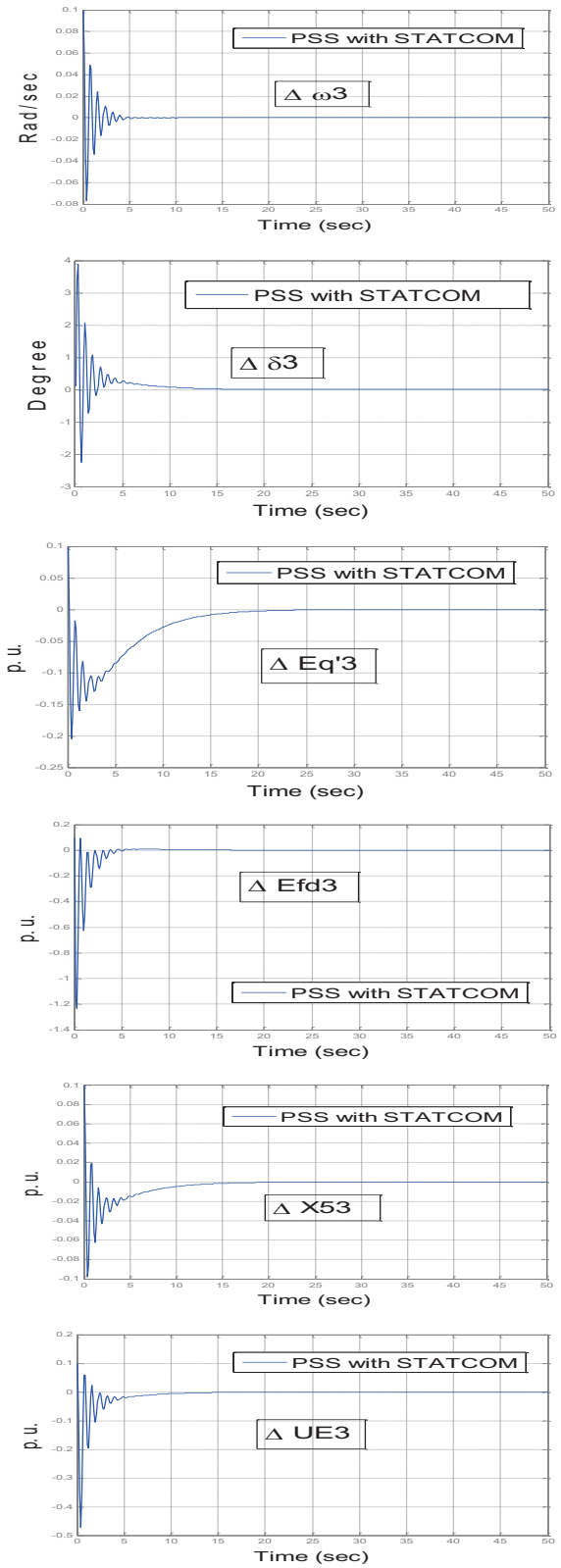


Figure. 9 Perturbation responses of system states for Case II for Six Area System with STATCOM supplements PSS (Area 3)

TABLE I. COMPARISON OF SYSTEM RESPONSE WITH ALL CASES FOR SIX AREA SYSTEM (AREA 3)

System	State	Over-shoot	Settling time	Eigenvalues
Case I (Previously tuned PSS)	$\Delta\omega 3$	0.05	6.10	-3.0591+14.3573i
	$\Delta\delta 3$	-6.0	6.50	-3.0591-14.3573i
	$\Delta Eq'3$	3.1	6.60	-12.9564
	$\Delta Efd3$	75.0	6.20	-0.7330 + 3.0407i
	$\Delta X53$	0.05	6.10	-0.7330 - 3.0407i
	$\Delta UE3$	2.40	6.50	-0.1227
Case I (Retuned PSS with KIM)	$\Delta\omega 3$	0.035	2.60	-5.3379 +11.8651i
	$\Delta\delta 3$	-4.0	2.50	-5.3379 -11.8651i
	$\Delta Eq'3$	1.80	2.70	-2.8189 + 4.9954i
	$\Delta Efd3$	50.0	2.80	-2.8189 - 4.9954i
	$\Delta X53$	0.035	2.60	-2.7471
	$\Delta UE3$	1.0	2.50	-0.1245
Case II PSS Fails to Stabilize even with GSA Driven KIM	$\Delta\omega 3$	---	---	-15.0770
	$\Delta\delta 3$	---	---	**0.2115 + 8.9028i**
	$\Delta Eq'3$	---	---	**0.2115 - 8.9028i**
	$\Delta Efd3$	---	---	-2.4103 + 0.8128i
	$\Delta X53$	---	---	-2.4103 - 0.8128i
	$\Delta UE3$	---	---	-0.1256
Case II STATCOM as supplement-ary controller with PSS	$\Delta\omega 3$	-0.08	5.0	-5787.400
	$\Delta\delta 3$	3.90	8.0	-0199.800
	$\Delta Eq'3$	-0.20	17.0	-0.900 + 7.8000i
	$\Delta Efd3$	-1.20	5.0	-0.900- 7.8000i
	$\Delta X53$	-0.10	12.0	-2.200 + 6.000i
	$\Delta UE3$	-0.46	8.0	-2.200 - 6.000i

VI. CONCLUSION

This paper presents an intelligent power oscillation damping control utilizing dynamical knowledge domain linked inference mechanism. The concept of controller sharing and shifting, to damp oscillations with dynamical change in operating condition, has been demonstrated. GSA has been used to develop the knowledge inference mechanism for controllers. The proposed controller shifting concept demonstrates that range of PSS at different operating condition is inadequate and change in operational shift will increase oscillations further, thus by retuning of PSS parameters the oscillations are quickly damped. The controller sharing concept demonstrates an effective role of STATCOM switching which is suitably connected in system at certain operating condition where PSS fails to stabilize. It is noticed that PSS with STATCOM enhances power modulation as desired, and thus stabilize the unstable system. The simulation results demonstrate that the over-shoot/ under-shoot along with settling time of the system state variables are greatly reduced.

VII. APPENDIX

All the constant K1 to K6 given in equation (16)-(19) can be calculated by following equations:

$$\begin{bmatrix} K_{1i1} \\ K_{2i1} \end{bmatrix} = \begin{bmatrix} 0 \\ i_{q1} \end{bmatrix} + \begin{bmatrix} F_{di1} & F_{qi1} \\ Y_{di1} & Y_{qi1} \end{bmatrix} \begin{bmatrix} (x_{qi1} - x'_{di1}) \\ e'_{qi1} + (x_{qi1} - x'_{di1}) * i_{d1} \end{bmatrix}$$

$$K_{3i1} = 1/[1 + (x_{di1} - x'_{di1}) * Y_{di1}]$$

$$K_{4i1} = (x_{di1} - x'_{di1}) * F_{di1}$$

$$\begin{bmatrix} K_{5i1} \\ K_{6i1} \end{bmatrix} = \begin{bmatrix} 0 \\ (v_{qi1} / v_{ti1}) \end{bmatrix} + \begin{bmatrix} F_{di1} & F_{qi1} \\ Y_{di1} & Y_{qi1} \end{bmatrix} \begin{bmatrix} -(x_{di1} * v_{qi1}) / v_{ti1} \\ (x_{qi1} * v_{di1}) / v_{ti1} \end{bmatrix}$$

In the above constants (K1 to K6), i represent the i^{th} generator and 1 is for area 1 and is extended for multi area multi generator system.

Data of Generator [21]:

Area 1 with PSS1 (G1):

Generator 1 (100 MW): M1=16.64 MJ/MVA; T1d0=5.6 sec; X1d=1.192; X1q=1.192; X1d1=0.1269; E1q1=1.0p.u.
Excitation System 2: K1a=18.5; T1a=0.2 sec;

Area 2 with PSS2 (G2):

Generator 2 (184 MW): M2=27.94 MJ/MVA; T2d0=3.3 sec; X2d=0.4993; X2q=0.4849; X2d1=0.0789; E2q1=1.0p.u.
Excitation System 2: K2a=18.5; T2a=0.2 sec;

Area 3 with PSS3 (G3):

Generator 3 (135 MW): M3=6.52 MJ/MVA; T3d0=3.5 sec; X3d=0.8667; X3q=0.5207; X3d1=0.2467; E3q1=1.0p.u.
Excitation System 2: K3a=40; T3a=0.060 sec;

Area 4 with PSS4 (G4):

Generator 4 (100 MW): M4=16.64 MJ/MVA; T4d0=5.6 sec; X4d=1.192; X4q=1.192; X4d1=0.1269; E4q1=1.0p.u.
Excitation System 2: K4a=18.5; T4a=0.2 sec;

Generator 5 (135 MW): M5=6.52; MJ/MVA; T5d0=3.5 sec; X5d=0.8667; X5q=0.5207; X5d1=0.2467; E5q1=1.0 p.u.
Excitation System 1: K5a=40; T5a=0.060 sec;

Generator 6 (140 MW):M6=16.1 MJ/MVA; T6d0=7.9 sec; X6d=1.54; X6q=1.49; X6d1=0.1060; E6q1=1.0p.u.
Excitation System 2: K6a=45; T6a=0.060 sec;

Area 3 and Area 4 connected with STATCOM:

Transformer: X_{te}=0.03;
Transmission line: X_e=0.3;
Operating conditions: V_E=1.0 p.u.; δ=40 degree;
DC Link Capacitor: C_{dc}=0.0005; V_{dc}=1.0 p.u.

Gravitational Search Algorithm Parameters:

No. of Populations: 70
No. of Iteration: 15
G0 (Gravitational constant)=100; α=20

REFERENCES

[1] Anderson, P.M, Fouad, A.A, "Power System Control and Stability", Iowa state university press,1977.

[2] Higorani, N.G., and Gyugyi, L., "Understanding FACTS: Concepts and Technology of Flexible AC Transmission Systems", IEEE Press, 1999.

[3] P. Kundur, "Power System Stability and Control", McGraw-Hill, 1994.

[4] Kundur P., Klein M., Rogers G., Zywno M., "Application of power system stabilizers for enhancement of overall system stability", IEEE Trans. Power Syst., 4, 2002, No. 2, 614-626.

[5] Abdel-Magid YL, Abido MA, Al-Baiyat S, Mantawy AH. "Simultaneous stabilization of multimachine power systems via genetic algorithms", IEEE Trans Power Sys 1999;14(4):1428–39.

[6] Abido MA, Abdel-Magid YL. "Robust design of electrical powerbased stabilizers using tabu search", IEEE Power Eng Society Summer Meeting, vol. 3, 15–19 July 2001, p. 1573–78.

[7] Eslami M., Shareef H., Mohamed A., "Tuning of Power System Stabilizers Using Particle Swarm Optimization with passive congregation", Inter. J. Phys. Sci.,5, 2010, No. 17, 2574-2589

[8] Eslami M., Shareef H., Mohamed A., Khajehzadeh m., "Damping of Power System Oscillations Using Genetic Algorithm and Particle Swarm Optimization", Inter. Rev. Electr. Eng., 5, 2010, No.6, 2745-2753.

[9] Wang H. F., Swift F. J., "A Unified Model for the Analysis of FACTS Devices in Damping Power System Oscillations. Part I: Single-Machine Infinite-Bus Power Systems", IEEE Trans. PWRS, 12, 1997, no. 2, 941–946.

[10] R. K. Pandey and N. K. Singh, "Small Signal Model for Analysis and Design of FACTS Controllers", IEEE PES General Meeting 2009.

[11] Eslami M., Shareef H., Mohamed A., Khajehzadeh M., "Particle Swarm Optimization for Simultaneous Tuning of Static Var Compensator and Power System Stabilizer", Przegląd Elektrotechniczny (Electr. Rev.) 87, 2011, No. 9a. 343-347.

[12] L. Khon and K. L. Lo, "Hybrid Micro-GA based FLCs for TCSC and UPFC in a Multi Machine Environment", Elec. Power Syst. Res. 2006, 76: 832-843.

[13] A. T. Al-Awami, Y.L. A. Magid and M.A. Abido, "A Particle Swarm-Based Approach of Power System Stability Enhancement with Unified Power Flow Controller", Elect. Pow. En. Syst., vol. 29:pp. 251-259, 2007.

[14] S. Jalilzadeh, H. Shayeghi, A. Safari and D. Masoomi, "Output Feedback UPFC Controller Design by Using Quantum Particle Swarm Optimization", IEEE Int. Conference ECTI-CON 2009, PATTAYA, THAILAND, pp. 28-31.

[15] Saied. Jalilzadeh and Amin Safari, "Design of State Feedback Damping Controller for the UPFC Using PSO technique", ICIS, IEEE International conference 2009, pp. 99-102.

[16] R. K. Pandey and N. K. Singh, "An Analytical Approach for Control Design of UPFC", Power System Technology and IEEE power India Conference, POWERCON Joint International Conference, 2008.

[17] R. K. Pandey "Analysis and Design of Multi-Stage LQR UPFC", Power, Control and Embedded System (ICPCES), 2010 International Conference IEEE, 2010.

[18] R.K. Pandey, Deepak K. Gupta,"PSS Tuning with Firefly Driven Knowledge Domain – A Smart Control Concept", In: IEEE TENCON, Macau, China; 1–4 November, 2015.

[19] Rajendra K.Pandey and Deepak K.Gupta, "Knowledge Domain States Mapping Concept for Controller Tuning in an Interconnected Power Network", Electrical Power and Energy Systems 80, 160–170, Elsevier, January 2016.

[20] Esmat Rashedi, Hossein Nezamabadi-pour and Saeid Saryazdi, "GSA: A Gravitational Search Algorithm", Information Sciences 179, 2009, 2232–2248.

[21] Yu YN, "Electric Power System Dynamics", Academic Press. 1983.

A method to update an ontology : simulation

Aly Ngoné NGOM[1], Yaya TRAORE[1,2], Papa Fary DIALLO[1], Fatou KAMARA-SANGARE[1] and Moussa LO[1]
[1]LANI, Gaston Berger University, Saint-Louis, Senegal
[2]LAMI, Ouagadougou University, Ouagadougou, Burkina
[1]{ngom.aly-ngone, diallo.papa-fary, fatou.kamara, moussa.lo}@ugb.edu.sn
[2]yaytra@yahoo.fr

Abstract— *In this paper, we present a method to add a new concept in a specific ontology (called basic ontology O_b). This method uses a reference ontology (called general ontology O_g) to add a new concept in a basic ontology. The reference ontology has the new concept and all the concepts of the basic ontology. The proposed method has three steps. Firstly, we use a semantic similarity measure in the reference ontology to find the most similar concept to the concept to add in the basic ontology. Secondly, we look for the right position of the concept to add relative to the most similar concept in the basic ontology. Finally, we insert the concept in the basic ontology by respecting its hierarchical structure. To illustrate our method, we use the whole WordNet[1] as the reference ontology and a branch of WordNet as basic ontology.*

Keywords: Ontologies, semantic similarity, add a new concept

1. Introduction

Ontologies are used as support for knowledge organisation by allowing users to annotate resources with regards to ontologies in a domain. However, the evolution of the domain reveals new concepts which do not exist in the ontology. There are three major approaches to add concepts in an ontology :

- one approach can be to use an expert (manual) to add these new concepts in the ontology ;
- another approach is automatic by using an algorithm to add these new concepts ;
- and a mixed approach combining the two previous approaches.

The methodology proposed in this paper is an automatic approach by using a reference ontology which is more general than the basic ontology. The reference ontology includes all concepts of the basic ontology and the new concept to be added. Our approach uses a semantic similarity measure to add this concept.
This paper continues by presenting our research context. Then, the third section explains the method that we propose to add a new concept in an ontology. The fourth section presents the simulation of our methodology by using WordNet. The fifth section presents the related works. We end with a conclusion and perspectives of this work.

[1]http ://wordnet.princeton.edu

2. Context

This paper is the continuation of the work presented in [1]. Indeed, we are using an ontology in a collaborative environment [12]. This collaborative environment allows users to annotate resources using a sociocultural ontology [12] or using tags. When they use tags it means they don't find any correspondence in the sociocultural ontology. The work done in [1] extracts all tags and applies a mining of frequent pattern. Thereby, the tags provided by the mining are candidated as new concepts of the sociocultural ontology. The interest of our work in this paper is to find the right position of these candidates in the sociocultal ontology without using any expert.

3. A method to add a new concept in an ontology

In this section, we propose a method to add a new concept C in a basic ontolgy O_b using a general ontology O_g. The ontology O_g includes, in its structural representation, all O_b's concepts and the new concept to add in O_b. Thus, this method has three steps :

- **step 1** : Find in O_g the concept C_{sim} in the basic ontology O_b which is the most similar to the concept C ;
- **step 2** : find the position of C relative to C_{sim} in O_g ;
- **step 3** : insertion of the concept C in the ontology O_b.

3.1 Find in O_g the concept C_{sim} the most similar to the concept C of the basic ontology O_b.

We use a general ontology O_g for looking for a concept semantically similar to the new concept C. As O_g contains all O_b's concepts and the new concept C, we use semantic similarity measure based on edges to find the concept C_{sim} the most similar to the concept C in O_g. The concept C_{sim} is also member of the ontology O_b. In [2], we have studied

semantic similarity measures based on edges and we have found that the measure of Zargayouna [3] presents a good correlation with the human judgement.
Zargayouna's measure is an improvement of Wu & Palmer 's measure [13]. It allows to assess the semantic similarity between two concepts favoring the similarity between concepts that are on the same hierarchy. We use this measure to assess the similarity between C and all O_b's concepts in O_g ontology. Algorithm 1 describes the method. In this

Algorithm 1: findConcept : Find in O_g the concept C_{sim} the most similar to the concept C of the basic ontology O_b.

input : C :concept; $stackO_b$: stack; O_g : ontology
output: C_{sim} : concept
variables : C_{sim}, A : Concept; $valco$, X : integer
$valco \leftarrow 0$
$C_{sim} \leftarrow$NULL
while *(!stackO_b)* **do**
  $A \leftarrow$depilate (stackO_b)
  $X \leftarrow$SIM (C, A)
  if *($X > valco$)* **then**
    $valco \leftarrow$X
    $C_{sim} \leftarrow$A
  end
end
return C_{sim}

algorithm, we have as input, the new concept C, the stack for all O_b'concepts storage named $stackO_b$ and the general ontology O_g. The algorithm depilates $stackO_b$'s concept and estimates the semantic similarity of them with the concept C in O_g. We get in output a O_b's concept named C_{sim} which is the more semantically similar to C. SIM is a function that implement Zargayouna's semantic similarity measure.

3.2 Find the position of C relative to C_{sim} in O_g.

In this step, we determine the position of the concept C relative to C_{sim} in O_g. The determination of the position allows to know if :

- C is the ancestor of C_{sim} in O_g ;
- C is descendant of C_{sim} in O_g ;
- C and C_{sim} are at the same level in O_g.

As O_g is more general than O_b, then the hierarchical structure of O_b will be respected in O_g.
For example, X is the new concept and Y is a concept of O_b. If X is ancestor of Y in O_g then X is the father of Y in O_b. If X is descendant of Y in O_g then X is the son of Y in O_b. Finally, if X and Y are in same level in O_g then X and Y are brothers in O_b.
The algorithm 2 is used to find the position of C relative

Algorithm 2: findConceptLevel : Position of C relative to C_{sim} in O_g.

input : C, C_{sim} : concept; O_g : ontology
output: integer
variables :$p1$, $p2$: integer
$p1 \leftarrow$depth(C,O_g)
$p2 \leftarrow$depth(C_{sim}, O_g)
return $p1 - p2$

to C_{sim} in the general ontology O_g. The function findConceptLevel takes as input the concepts C and C_{sim} and the ontology O_g and returns an integer value. If the returned value is :

- positive then C is the son of C_{sim} in O_b ;
- equal to 0 then C and C_{sim} are brothers in O_b.
- negative then C is the father of C_{sim} in O_b.

3.3 Insertion of the concept C in the ontology O_b.

After we have found the concept C_{sim} the more semantically similar to C in O_g and its position relative to C, the last step is to insert C in the basic ontology O_b. This operation is realised by the algorithm 3. This algorithm allows to insert a

Algorithm 3: insertConcept : Insertion of the concept C in the ontology O_b.

input : C,C_{sim} :concept ; $stackO_b$: stack; O_g, O_b : ontology
output: O_b : ontology
variables : $C2$: concept; $value$, X : integer; $stackofFather$: stack of Concepts ;
$C_{sim} \leftarrow$findConcept(C, stackO_b, O_b)
$value \leftarrow$findConceptLevel(C,Csim,O_b)
if *($value < 0$)* **then**
  $ancestor(C, C_{sim})$
  $ancestor(Thing, C)$
end
if $value > 0$ **then**
  $descendant(C, C_{sim})$
end
if *($value = 0$)* **then**
  $stackofFather \leftarrow$stackFather(Csim,O_b)
  $C2 \leftarrow$findConcept (C, stackofFather, O_g)
  $descendant(C, C2)$
end
return O_b

new concept C in the ontology O_b respecting its hierarchical structure. In this algorithm, we call the algorithms 1 and 2 to find C_{sim} and its position relative to C in the ontology O_g. We have three kind of insertion :

- if C is an ancestor of C_{sim} in O_g, we insert it like a father of C_{sim} in O_b using the function ancestor() ;
- if C is a descendant of C_{sim} in O_g, we insert it as a son of C_{sim} in O_b using the function descendant() ;
- otherwise, if C is at the same position of the concept C_{sim} in O_g, we insert it in O_b like a brother of C_{sim}. To insert C as a "brother" of C_{sim}, we first stack all the "father" of C_{sim} in O_b, then we use the function findConcept() to find the more similar concept to C in the stack using the ontology O_g and finally when we have found the concept, we use the function descendant() to insert C like a son of C_{sim}'s "father" (then C is the "brother" of C_{sim}) in O_b.

To finish, the Algorithm 3 returns the ontology O_b updated.

4. Simulation

In this section, we do a simulation of the proposed method. In the simulation, we use WordNet as the general ontology(O_g). We use a subontology of WordNet as the basic ontology (O_b) and we add to it some concepts by using the proposed method. In order to evaluate the semantic similarity between the news concepts and O_b's concepts, we use WordNet-Similarity [11]. The figure 1 is the subontology extracted from WordNet.

Given two concepts "boat" and "knife". Our aim is to add

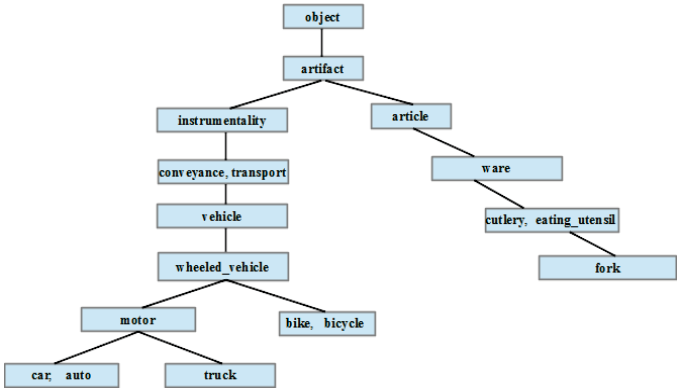


Fig. 1: The basic ontology extracted from WordNet.

the two concepts in the basic ontology. The table 1 presents the result of our method when we add "boat".

In table 1, we note that "vehicle" is more similar than "boat". When we evaluate the position of "boat" relative to vehicle in WordNet, we get a positive value (depth(boat) - depth(vehicle) = 3). Therefore, "boat" will be a descendant of "vehicle" in O_b. The figure 2 shows the result of the insertion.

By reproducing the same procedure with the concept "knife", we obtain as results the table 2 and the figure 3.

Since the result is positive, then "knife" is descendant of "cutlery" and "eating_utensil".

O_b's concepts	Result of similarity between "boat" and O_b's concepts in WordNet	Position of "boat" relative to the more similar concept in WordNet	Relation between the two concepts in O_b
object	0.57		
artifact	0.75		
instrumentality	0.82		
article	0.74		
conveyance	0.8		
transport	0.8		
ware	0.8		
vehicle	0.86	3	boat is a descendant of vehicle
wheeled_vehicle	0.16		
cutlery	0.15		
eating_utensil	0.15		
motor	0.05		
bike	0.116		
bicycle	0.116		
fork	0.1		
truck	0.1		
car	0.116		
auto	0.07		

Table 1: Result of insertion of "boat" in O_g.

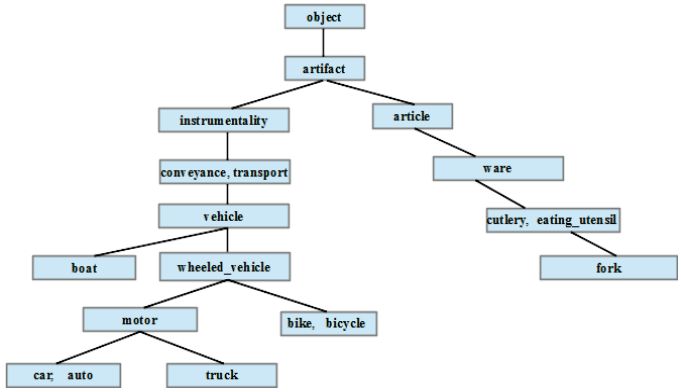


Fig. 2: The basic ontology after adding "boat".

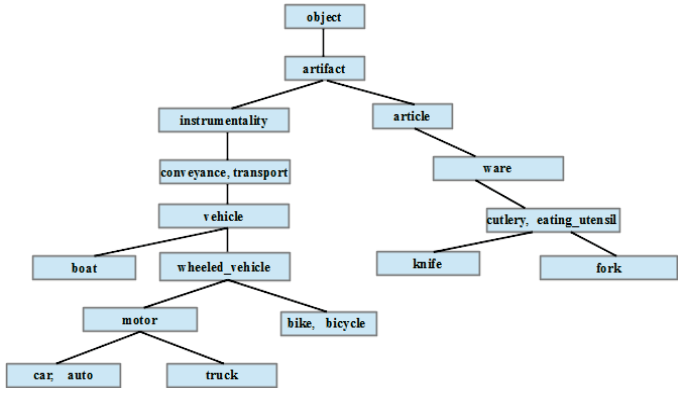


Fig. 3: The basic ontology after adding "boat" and "knife".

O_b's concepts	Result of similarity between "knife" and O_b's concepts in WordNet	Position of "boat" relative to the more similar concept in WordNet	Relation between the two concepts in O_b
object	0.53		
artifact	0.7		
instrumentality	0.78		
article	0.06		
conveyance	0.08		
transport	0.09		
ware	0.04		
vehicle	0.06		
wheeled_vehicle	0.06		
cutlery	0.92	2	knife is a descendant of cutlery
eating_utensil	0.92	2	knife is a descendant of eating_utensil
motor	0.09		
bike	0.04		
bicycle	0.04		
fork	0.12		
truck	0.04		
car	0.04		
auto	0.03		
boat	0.04		

Table 2: Result of insertion of "knife" in O_g.

5. Related Works

The aim of our work is to add a new concept in an ontology. We can distinguish two important approaches of changes management in ontologies [4] : ontology versioning and ontology evolution. Ontology versioning consists in building and managing different versions of an ontology while allowing access to these versions. Ontology evolution, in turn, is a process that changes the ontology while keeping it consistent. In this approach, there are three major operations :

- conceptual change that represents changes in conceptualizing ontology ;
- the specification change that only affects specific parts of the conceptualization such as changing properties of a concept ;
- the representation change in conceptualizing such as the change of language representation of ontology.

Our work is part of ontology evolution and particularly, the conceptual changes operation. In this domain, Flahive and colleagues propose to extract and extend a subontology and merged it with another one [5][6]. They present two methods : the minimum extraction and the maximum extraction. The minimum extraction consists in selecting concepts in an ontology and extracting them. The maximum extraction marks the concepts of the ontology before extracting them. There are three kinds of mark : "selected", "unselected" and "undecided". In the maximum extration, the "selected" and "undecided" concepts are extracted. Nevertheless, the two methodologies present some weakness. Indeed, after ontologies are merged, some concepts are deleted and consequently information is lost. In addition, contrary to our approach, they don't present a method to link the descendant of a deleted concept to the ontology.

There are some tools created for ontology evolution. We can mention GLUE [7], OntoMerge [10] and Chimaera [8]. Chimaera [9] allows merging and ontology management based on semantic integration. Its operation is based on multiple levels as the comparison of classes and diagrams to the ontological fusion through interference established between the two sources ontologies. OntoMerge [10] also conducts semantic integration and makes the ontological fusion with an inference system. GLUE, meanwhile, uses learning techniques (Machine Learning) to the discovery of mapping between source ontology concepts.

6. Conclusion and perspectives

In this paper, we have proposed a method for updating ontology. This method adds new concepts in a basic ontology using a general ontology. For experimentation, we have used a WordNet's subontology as basic ontology and WordNet as general ontology. The proposed method has three major steps :

- find a concept of the basic ontology that is more similar to the new concept based on general ontology ;
- determine the new concept's position relative to the concept found in the first step and in the last step ;
- and insert the new concept in the basic ontology.

As future works, we plan to use this method to add concepts in a sociocultural ontology [12]. In another perspective, we will verify the semantic relevance of the added concept related to its neighbors in the ontology. The neighborhood of the concept is composed of the concepts which are similar to it in the ontology. This step will allow us to discuss the relevance of this added concept in the ontology.

References

[1] Y. TRAORE, S. MALO, C. T. DIOP, M. LO, O. STANISLAS. *Approche de découverte de nouvelles catégories dans un wiki sémantique basée sur les motifs fréquents.*, IC2015 Rennes, France. 2015, pp. 129 - 134.

[2] A. N. NGOM. *Étude des mesures de similarité sémantique basée sur les arcs.*, CORIA, Paris, France 2015, pp 535 - 544.

[3] H. ZARGAYOUNA, S. SALOTTI. *Mesure de similarité sémantique pour l'indexation de documents semi-structurés.* In 12ème Atelier de Raisonnement à Partir de Cas, Mars 2004.

[4] B. YILDIZ. *Ontology Evolution and Versioning : the state of the art.* Vienna University of Technology Institute of Software Technology Interactive Systems (ISIS) Asgaard-TR-2006-3. 2006 p. 28.

[5] C. FLAHIVE, D. TANIAR,J.W. RAHAYU. *Ontology as a Service (OaaS) : extending sub-ontologies on the cloud.* Concurrency Compuat. . IPrat, Exper. 2015 ; 27 pp:2028 - 2040.

[6] N.F. NOY AND M.C. MUSEN. *PROMPT : Algorithm and tool for automated ontology merging and alignment.* Proc 17th Natl. Conf. On artificial Inteligent (AAAI2000), Austin, Texas, USA 2000, pp. 450 - 455.

[7] A.H. DOAN, J. MADHAVAN, P. DOMINGOS, A.Y. HALEVY. *Learning to map between ontologies on the semantic web.* 11th International Conference on World Wide Web. 2002. pp. 662 - 673.

[8] D.L. MCGUINNESS, R. FIKES, J. RICE, S. WILDER. (2000). *An environment for merging and testing large ontologies.* Proceedings of the 7th International Conference Principles of Knowledge Representation and Reasoning, (KR'2000), Colorado, USA, 2000 ; pp. 483 - 493.

[9] N. F. NOY. *Semantic integration : C survey of ontology-based approaches Principles and Practice of Semantic Web Reasoning.* Lecture Notes in Computer Science Volume 4187, 2004, pp 164-178.

[10] D.DOU, D.V. MCDERMOTT, P. QI. *Ontology translation on the semantic web.*. journal on Data Semantics 2005 ; 2 pp:35 - 57

[11] T. PEDERSEN, S. PATWARDHAN, J. MICHELIZZI. *WordNet::Similarity - Measuring the Relatedness of Concepts.* In Proceedings of the Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics. Boston, MA. 2004. pp. 38 - 41.

[12] P.F. DIALLO, O. CORBY, M. LO, I. MIRBEL, S.M. NDIAYE. *Sociocultural ontology : upper-level and domain ontologies.* In Acts JFO, Tunisie 2014, pp : 15 - 27.

[13] Z. WU, M. PALMER, *Verbs semantics and lexical selection.* In U. A. f. C. L. Stroudsburg, PA (ed.), In Proceedings of the 32nd annual meeting on ACL, volume 2 de ACL '94, p. pp : 133 - 138, 1994.

Cloud Security Management Model based on mobile agents and web services interaction

Abir KHALDI[1], Kamel KAROUI[1], Henda BEN GHEZALA[1]
[1] RIADI Laboratory ENSI, University of Manouba, Manouba, Tunisia

Abstract- *Security is one of the most important issue faced by the cloud adoption. Therefore, cloud actors such as customers, provider, business partners, and auditors are asking for major security controls and measures to be set in order to supervise and protect cloud assets and services. In fact, the security management is a very complex task specially in cloud environment because of its multi-layers services and multi-tenancy. In this paper, we propose a universal cloud security management model to cover all cloud services. This model is based on four phases: the perception, the detection, the reparation and the evaluation. The last phase offers a security assessment for each cloud service and also for cloud hypervisor to evaluate cloud service based on more than one security assessment indicator. The cloud security management model profits of the advantages of mobile agent and web service interaction.*

Keywords: cloud security, SIEM, mobile agent, web service

1. Introduction .

NIST [1] defines three cloud services models: Software as a service (SaaS), Platform as a service (PaaS) and Infrastructure as a service (IaaS). Those cloud services may suffer from several vulnerabilities which are due to design, programming, or configuration errors. Such vulnerabilities can be exploited by malicious users to succeed their attacks [2][3][4].

In fact based on many studies [5][6][7][8][9], cloud adoption willingness was tightly related to security concerns. Therefore cloud services need to guarantee more security in order to sell better.

The idea of this work is to propose a universal cloud security management model for all cloud services model. This model collects and correlates cloud services event log to detect vulnerabilities and/or attacks in order to repair any anomaly detected. An evaluation step to assess cloud service security is an integrated part of this model. It gives more than one security assessment indicator to measure the cloud service security level. To ensure a dynamic model, we propose to introduce a smart autonomous mobile agent interacting with web service to correlate event log between different cloud assets and to automatically repair anomaly if detected.

The remainder of the paper is organized as follows. In Section 2, we give an overview of related work in cloud security management in the cloud. Section 3 presents the proposed cloud security management model. Furthermore, we define the different model components then we describe the four phases of the model. Finally, Section 5 concludes our paper and describes our future work.

2. Related Work

Many studies focus on the security management in the cloud environment, as a requirement for cloud business evolution.

In [10] an automated evaluation of cloud security mechanisms and their efficiency is proposed. The access control and the intrusion detection systems are the main objectives of this research. This approach concerns only the cloud infrastructure.

The Cloud Security Alliance (CSA) [11] professionals and researchers presented the concept of Security-as-a-Service (SECaaS) to cloud services. They developed a set of requirements, and discussed implementation considerations and concerns. However, the provided recommendations did not detailed a specific model covering all cloud services.

Niekerk et al [12] proposes a model to integrate traditional security solutions into a cloud infrastructure. In fact, their model presents a high level description and does not provide any details on the implementation and evaluation of the security in the cloud infrastructure, platform or software layer.

Hussain et al [13] introduced SECaaS using service oriented architecture (SOA) to allow cloud customers to have more control over hosted services. A user-centric approach was employed to allow users to choose security services and monitor the status of their applications and data in the cloud environment. However, their architecture is only focusing on the access control settings and some security settings in the chosen cloud service model (IaaS, PaaS, or SaaS).

In [14], the authors studied security controls recommended by standards such as ISO/IEC 27001 and NIST SP 800-35. They noticed that 30% of the controls can be automated. They introduced a security information and event management (SIEM) framework to automate these security controls in this work. But, they did not consider the application of their framework on the multi-layer/multi-tenancy architecture of a cloud computing environment.

3. Proposed Cloud Security Management Model (*CSMM*)

In this section we present the CSMM components and we describe the different CSMM phases as shown in figure 1.

3.1 *CSMM components*

The CSMM is composed of 4 principal components:

- ✓ Sensors : The CSMM exploit sensors outputs to recognize what can occur in different cloud services models (IaaS/PasS/SaaS). We deploy two types of sensors:
  - Service Sensors : sensors deployed in different cloud service (SaaS, PaaS, IaaS) can be a log manager, an IDS/IPS, Web application Firewall (WAF), etc.
  - Hypervisor Sensors: The hypervisor is a critical component for cloud environment. Hypervisor sensor can be a log manager sensor to collect all VMs events and/or an NIDS/NIPS to detect and/or to prevent cloud attacks.

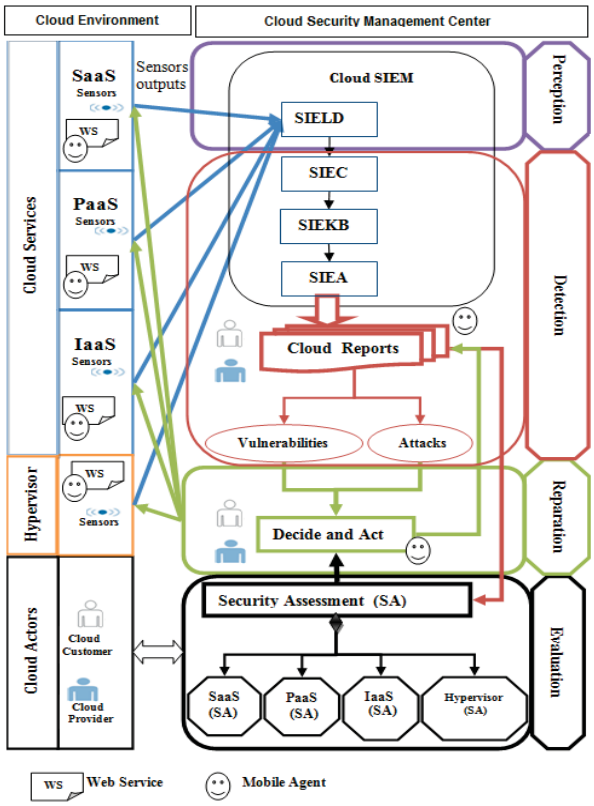


Fig. 1. Cloud Security Management Model (*CSMM*)

- ✓ Mobile agent (*MA*) [15]: it is a smart mobile code which can migrate with a base rules to detect and repair cloud anomalies (vulnerabilities, attacks).
- ✓ Web service (*WS*) [16] : it is the intermediate between sensors outputs databases and the mobile agent. Mobile agent and web service interaction [17] helps to secure cloud assets and ensure a rapid and interoperable communication.
- ✓ Cloud actors : it can be a cloud customer or a cloud provider involved in the cloud security management model.

3.2 CSMM phases

The *CSMM* is based on four phases :

Phase 1 is the Perception phase. It constitutes the first step of the cloud SIEM (Security Information end Event Management). This step named SIELD (A Security Information and event log and database) considers the events and logs database (ELD) as a repository for all events and logs sent by the different cloud sensors. It is updated in real time and has a mirrored ELD backup as a contingency in case of failure**.**

Phase 2 is the Detection. It is composed of:

- A Security Information and event Correlation (SIEC) module: The correlation is a key step as it is used to detect events not previously noticed. It uses the information stored in the SIELD in order to provide meaningful results. The correlation results are evaluated to identify relationships and detect threats.
- A Security Information and event knowledge base (SIEKB) module: The knowledge base (KB) is an online known threat centralized repository for cloud customer infrastructure, platforms and software services. It contains symptoms that match certain event(s) along with the recommended counter measures and/or responses.
- A Security Information and event Analysis (SIEA) module: The security information and event analysis (SIEA) module allows cloud security analysts to perform advanced research on events. Some events need further explanation and investigation to provide additional details.

The cloud reports are XML documents which contains the results of the Cloud SIEM. We create a cloud report for each cloud service and an hypervisor cloud report.

Based on the cloud reports, we can detect if there is any vulnerabilities and /or attacks that should be repaired. In a previous work [18], we design a framework to detect a distributed cloud attacks in hybrid cloud based on MA/WS interaction.

Phase 3 is the Reparation. In fact, we propose two solutions to repair an existing anomaly: if the anomaly and its repair mode is known by the mobile agent, it can be automatically repaired. If it is not, the cloud provider and /or the cloud customer should decide and act to resolve problem. When the anomaly is repaired, it should be mentioned in the cloud report. A load balancing technique is proposed in our

previous work [19] to improve cloud services availability in cloud environment.

Phase 4 is the Evaluation. We exploit the cloud reports results to assess cloud security service. Therefore we will adopt the chen's et al [17] threat evaluation model for cloud service. After evaluation, the security assessment can help cloud actors to decide to make more security controls in their cloud services.

4. Summary and Future Work

In this work, we propose a universal *CSMM* to enhance security in cloud environment. The use of mobile agent instead of client/server model decreases the cloud traffic and distributes the processing charge between virtual machines. It gives *CSMM* an autonomous and dynamic aspect by repairing anomaly automatically within the smart mobile agent. In the next work, we are going to develop the evaluation phase by proposing a quantitative approach to measure the cloud security situational awareness.

5. References

[1] Mell, P. &Grance, T., 2011, "The NIST Definition of Cloud Computing", NIST Special Publication 800-145.

[2] R. Buyya, R. Ranjan, R. Calheiros. InterCloud: Scaling of Applications across multiple Cloud Computing Environments. In Proc. of the 10th Int. Conf. on Algorithms and Architectures for Parallel Processing, 2010.

[3] S. Roschke, F. Cheng, and C. Meinel, Intrusion detection in the Cloud, In Proc. of the 8th IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, 2009, pp. 729–734.

[4] N. Gustavo, C. Miguel. Anomaly-based intrusion detection in software as a service. In Proc. of the Dependable Systems and Networks Workshops, 2011, pp. 19–24.

[5] Cloud Industry Forum, "Cloud UK The Normalization of Cloud in a Hybrid IT Market UK Cloud Adoption Snapshot & Trends for 2015, http://itsmf.cz/wpcontent/ uploads/2014/09/CIF_White_Paper_Normalisation_of_Cloud_Zyn Branded.pdf, retrieved 2015-02-16

[6] Tsai, Chang-Lung, et al. "Information security issue of enterprises adopting the application of cloud computing." Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on. IEEE, 2010.

[7] Jensen, Meiko, et al. "On technical security issues in cloud computing."Cloud Computing, 2009. CLOUD'09. IEEE International Conference on. IEEE, 2009.

[8] So, Kuyoro. "Cloud computing security issues and challenges." International Journal of Computer Networks 3.5 (2011).

[9] Subashini, Subashini, and Veeraruna Kavitha. "A survey on security issues in service delivery models of cloud computing." Journal of network and computer applications 34.1 (2011): 1-11.

[10] T. Probst, E. Alata, M. Kaaniche, V. Nicomette, & Y. Deswarte, "An Approach for Security Evaluation and Analysis in Cloud Computing," Safecomp, France, September 2013.

[11] Cloud Security Alliance, "Security Guidance for Critical Areas of Focus in Cloud Computing v3.0," Cloud Security Alliance, https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf, 2011.

[12] B. Niekerk, & P. Jacobs, "Cloud-based security mechanisms for critical information infrastructure protection," IEEE International Conference on in Adaptive Science and Technology, pp. 1-4, November 2013.

[13] M. Hussain, & H. Abdulsalam, "SECaaS: security as a service for cloudbased applications,"ACM Conference on e-Services and e-Systems, Kuwait, April 2011.

[14] R. Montesino, S. Fenz, & W. Baluja, "SIEM-based framework for security controls automation," Information Management & Computer Security, Vol. 20, No. 4, pp. 248-263. 2012

[15] D. Lange, M. Oshima, "Seven Good Reasons for Mobile Agents", 1999. Communications of the ACM Issue.

[16] Web Service Activity Proposal, 2000.White paper. Ben Ftima F., Karoui, K.,"Interaction Mobile Agents - Web Services", Second Edition, pp. 717-725 edited by Encyclopedia of Multimedia Technology and Networking.

[17] Chen XZ etc., "Quantitative hierarchical threat evaluation model for network security", *Journal of Software*, 2006,17(4): pp.885-897.

[18] Khaldi, Abir, Kamel Karoui, and Henda Ben Ghezala. "Framework to detect and repair distributed intrusions based on mobile agent in hybrid cloud."Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science,Computer Engineering and Applied Computing (WorldComp), 2014.

[19] Khaldi, Abir, Kamel Karoui, and Henda Ben Ghezala. " Intra-cloud and inter-cloud Load balancing based on interaction between mobile agent and web service" PDPTA 2015.

Defining Adaptive Whitelists by Using Clustering Techniques, a Security Application to Prevent Toll Fraud in VoIP Networks

Santiago Israel Monterrosa Reyes
Facultad de Ingeniería
Universidad Anáhuac del Sur
Ciudad de México, México

Gerardo Reyes Salgado
Centro Nacional de Investigación y Desarrollo (CENIDET)
CIICAp, UAEM
Cuernavaca, Mor., México
greyes@cenidet.edu.mx

Joaquín Pérez Ortega.
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Cuernavaca, Mor., México
jpo_cenidet@yahoo.com.mx

***Abstract*— One strategy to prevent telephone fraud used by telephone operators is the so called White list which is a set of international destinations where the user is allowed to call, it is by default static and is changed manually by request of the user. This paper describes the creation of adaptive White list that will change based on the VoIP traffic behavior of corporate customers, starting with a list of countries based on the frequency of calls to every destination per customer, so an easy way to create a small set of Whitelists rather than a White list per customer, but equally effective to block fraud calls, is proposed. As well the Whitelists are adaptive so they are changing according to user group's behavior over the time. The Weka's *SimpleKMeans* method is used to cluster international destinations where the customers use to call on a six month period. We describe briefly the K-means method and a way to measure its effectiveness to generate clusters of destinations per customer, the validation method includes the EM algorithm which is briefly described as well. The generated clusters are tough to feed a machine learning system used to detect toll frauds to be developed as a next phase.**

Keywords*— machine learning, clustering, data mining, toll fraud, VoIP

I. Introduction

It is very common that current IP telephony operators must implement security best practices abroad its telecom infrastructure, from the core network servers to endpoints like softphones, smartphones, IP phones, and even ATA devices. Not to say about network equipment for switching and routing voice and video traffic all through the IP telephony network is also currently protected with traditional security techniques, as long as much of these infrastructures are based on IP technologies hence targeted by hackers and phreakers to commit toll fraud, even more currently telephone service is a software application and as such is exposed to a great variety of attacks.

This article is the initial phase of a project to develop a toll fraud detection algorithm that learns the behavior of telephone traffic coming from corporate users to an IP Telephony network, in the rest of this work we refer to users meaning corporate users or corporate customers served by a VoIP operator. This initial phase describes the White list generation process, the White list consists of a list of international destinations where every user is allowed to call; whereas the generation process includes the design of experiments to create White lists taking a history of six months behind of traffic behavior coming from users; it is expected that one user has a common profile with other users, but also the traffic behavior will change over the time in such a way that every periodic interval of time a new White list is generated defining this way an adaptive White list for a user that is common to other users in some sense. The goal is to obtain every time a new White list that is, as much as possible, more restricted than the previous one, based on the behavior of every user of the telephone operator referred to a traffic history of six months behind. The results are to be used as inputs of a machine learning system to detect frauds in a shorter time and very close to the beginning of the attack in regards to other fraud detection techniques based on CDRs.

K-means Method

It is a simple but widely used method to cluster unsupervised data, which means there is no clusters previously known and K-means algorithm tries to find out the optimal clusters, in certain sense, based on the Euclidean distance of the centroid to every point of the cluster.

An arbitrary number of k clusters is initially set along with the k centers of every cluster determined randomly. Every element or point finds out the closer center to it, hence the cluster it belongs to, so every center belongs to a set of points. Afterwards every center finds out the centroid of the cluster it belongs to. Then every centroid becomes in the center of the cluster, the sequence is the repeated till the end.

A formal description of K-means method is described in [13]. Trying to determine the optimal number of clusters is not easy; in a clustering process it commonly means that the objective is to minimize the Schwartz criteria [8].

One way to validate the K-means method is to contrast it against a hierarchical process.

Hierarchical clustering

In the hierarchical method is possible to start with *k* unknown, the process considers every point as its own cluster, and then it looks for similar pairs of points, in a certain sense. Then it clusters the similar points into a father cluster, then repeat the process until a dendogram or spanning tree diagram is obtained as shown in Figure 1.

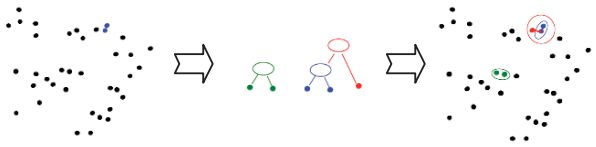


Fig. 1. Hierarchical clustering

Some techniques are incremental; two of them are Cobweb and Classit. There is a statistical technique based on a mixture of different probabilistic distributions, one per cluster. It assigns instances to a class in a probabilistic way, so different to K-means, it doesn't create disjoint clusters [3].

Clustering based on probabilities

The goal here is to find out the most probable set of clusters given the data. As there is no evidence to assign one instance to one category or another, an instance has certain probability to belong to one cluster or another. The foundation is a model called finite mixtures. A mixture is a set of *k* probability distributions representing *k* clusters ruling the values of the attributes for the members of that cluster.
Every cluster has a different distribution, any particular instance can belong to one cluster or another but is not known exactly to what. The clusters are unequally probable, and there is a probability distribution reflecting its relative populations. Every cluster is defined by the three parameters Probability **p**, mean **μ**, and standard deviation **σ** [3].

EM Algorithm (Expectation Maximization)

This is the algorithm against K-means will be validated for the given dataset.
It is not known the distribution of a particular instance belongs to, neither the five parameters of the mixture. The algorithm adopts the K-means algorithm and iterates it.
It starts guessing the value of the five parameters and are taken to calculate the probabilities of the cluster for every instance, in time the probabilities are taken to calculate the new five parameters and repeat it. This last calculation is the "maximization" of the likelihood of the distributions given the available data.
Unlike to K-means, EM algorithm stops when the instance of a class doesn't change from one iteration to another, the algorithm converges to a fixed point that actually never reaches. There is a way to get a measure of the "goodness" of clustering based on the probability of an individual instance times the probability of the other individual instances [3]

II. Related Works

The experience shows that fraudsters use every time more sophisticated techniques but crafted based on simplicity, so machine learning and data mining techniques have shown to be effective to mitigate toll fraud since late 90s [1]. An explanation of tools to build a fraud detection system as well as main components of it is given in [14-15]. As Whitelisting is a technique widely used in perimeter security devices, as intrusion prevention/detection systems, a proposal based on tracking deviations from interacting protocols using state machines in order to detect intrusions on VoIP systems is described in [2]. A review of Data Mining frameworks for Intrusion Detection including Clustering and Hierarchical algorithms is in [5]. A comprehensive description of several methods and techniques to detect and prevent fraud in VoIP Networks, telecommunications and other areas can be found in [11-12]. In general fraud detection systems are based on two categories, one is called *absolute analysis* based on the calling activity models of fraudulent behavior and normal behavior; and the other one is called *differential analysis* focused on detecting sudden changes in behavior, both approaches are detailed in [4], it describes a great set of techniques to implement analysis typically using probabilistic models, neural networks, or rule based systems. In [6] two techniques, neural networks and rule-based, are used to implement a fraud detection tool that profiles subscribers and network traffic, one application of these techniques is described in [7]. A recent work using SOM, K-means, and EM algorithms to detect toll fraud in VoIP networks by profiling user's behavior and setting up thresholds is described in [9-10].

III. Problem Description and Propoused Solution

Basically it consist of the abuse of telephony service by users that find the way to make calls over lines at no cost, either they do not lease the service or the lines were assigned to them by an organization as they are users as well, at the end the fraudster won't pay for the abused service, instead a third party will be charged, usually the telephone service operator. Telephone fraud is usually committed to make calls to either international destinations or cell phones or satellite destinations, as these destinations tag the straight way to the money that fraudsters normally follow, the bigger is the volume of calls the better is the business of the crooks.

For IP telephony operator the problem is bigger as its technology is mostly based on the SIP protocol that inherits the security issues of a traditional IP data network. As a reference, tele-phone operators lose around four billion dollar every year because of telephone fraud in the in-come shared by international traffic [2]. As well the operator loses reputation and confidence from its customers. Especially corporate customers stop seeing its telephone operator as a secure service provider. This situation encourages the development of fraud detection for IP telephony service

systems, although sophisticated fraud detection tools have been created and operators spend millions of dollar to prevent telephone frauds effectively, the criminal processes are adapted rapidly, hence the fraud detection tools must be adapted quickly as well so this is the sense of the present effort. Figure 2 shows a graphic representation of the problem for a VoIP operator.

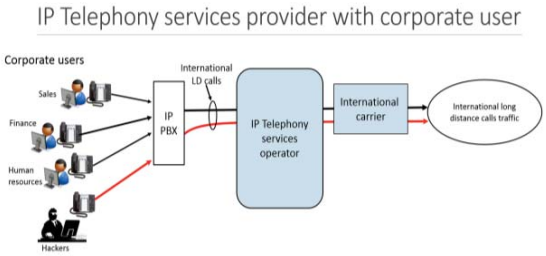


Fig. 2. IP Telephony Service Provider

A. Generation Adaptive White List by Clustering

The concept of Whitelist is based on a security strategy that states "every access is restricted except the ones that are explicitly allowed", the way it works for a telecom operator is to control calls directed to long distance international destinations in order to prevent telephone fraud. In terms of the telephony exchange platform some rules are implemented to specify ranges of telephone numbers assigned to a corporate customer and mapped to the list of allowed destinations to this customer; so every call going to an international destination originated by a corporate user is verified by originating number and destination, if it satisfies the criteria defined by the rule then it is routed to an international carrier switch to reach its destination otherwise the call is dropped.

The operator subject of the present work handles a single default White list composed by a list of 27 destinations determined by the historically more frequently called destinations measured in any given time several years ago. This is a static list unless the user requests to open more destinations, so over the time it is expected the list change defining somehow the individual behavior of the customer consumption of calls to international destinations.

The behavior of an individual White list may leave open the door to fraudsters, as a requested new open destination could be used only a few times by the user, afterwards the destination is not closed because the list is static, so once open this destination might be targeted by an attacker easily.

We will investigate if the White list of every user has to do with the behavior of other users in relation to the destinations they use to call. So we see if there is a convenience to have clusters of users, every cluster defining a new White list for the user belonging to the cluster.

Two benefits will have this technique, one is to have few White lists to implement them easily in a telephone platform instead of one for every user, and the other one benefit is to get a changing White list according to cluster's behavior over the time, instead of having a static list changed manually every time a fraud incident comes up, or a user gets out or in of the platform.

The results of K-means will be compared against the results of running the EM algorithm for the same set of data, a brief explanation will support this comparison as a method to validate the White list generation process proposed.

B. Data Preparation

The data to analyze were extracted from the so called CDR (Call Data Register) that records every detail of a call: originating number, destination number, start date and time of the call, as well as, termination date and time of the call, duration of the call. Some other specific data are included in a CDR according to the specific needs of the operator; the operator generates CDRs from three different platforms, every platform for specific telephone service for corporate customers: hosted PBX, SIP Trunking, and SS7 based telephony. The main interest of the operator to generate CDRs is the billing process, the CDRs coming from every different platform have different layouts, so they are homologated in order to be processed for billing, the homologation process is called mediation and is shown in Figure 3 below.

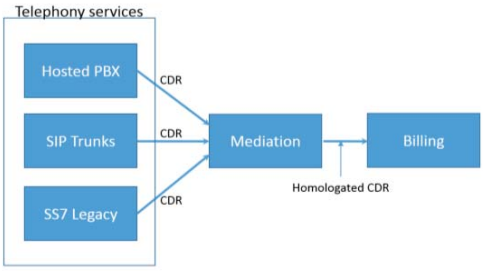


Fig. 3. Mediation Process

The CDRs can be also used to generate traffic reports showing the behavior of every user, some simple traffic patterns can be defined to monitor continuously the traffic behavior and detect potential fraud calls. For instance, the current call volume from a specific user is compared against its historic traffic three months behind, to the same destination and at the same time of the day. The operator uses the standard deviation compared against the history of traffic two months behind as a metric to detect potential fraud which is also simple but effective when an operations engineer with some good knowledge about traffic behavior is monitoring and analyzing it carefully.

From different analysis of the traffic behavior, it has been estimated a volume of traffic at peak hours around 15000 calls per second and total monthly traffic of 50 million calls. In order to get an idea of the difficulty to detect fraud calls, the irregular traffic can consist of three or four calls per hour from midnight to eight in the morning. Afterwards this traffic may growth to hundreds of calls per hour at 10 am, and later around noon, peak hours, it becomes at higher volume say thousands of calls per hour, that are difficult to identify as potential fraud because its behavior is pretty similar to the

regular traffic. It is shown in figure 4 below where fraud 1 is easy to identify and fraud 2 is hard to identify.

All the information about the volume of calls and minutes done by users to international destinations during the second semester of 2014 were extracted from the homologated CDR data-base. Yet some clean up must be done to the output data,

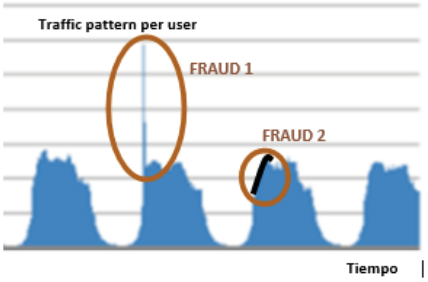


Fig. 4 Fraud pattern hard to detect

as every telephony platform handle names of cities and countries differently. Finally a list with the following fields was obtained

<customer_name,destination,call_volume,duration>

From this list a total universe of 550 users was obtained, from there 69 have all of the world-wide destinations open.

At this point is where it should be decided if is better take call volume, duration in minutes or both to profile the behavior of user's traffic, by reviewing the history of fraud incidents looks like the call volume is the indicator that helped to detect the incident and the take action. So for this analysis, a more simplified list is used

<customer_name,destination,call_volume>

In order to process the data appropriately the fields customer_name and destination are alphanumeric, while call_volume is binary, where 1 means the user call to de destination, and 0 means the corporate didn't call to the destination in the period of six months. Although certain granularity is missed taking only the fact that the corporate called to the destination instead of taking the number of calls, it is good enough to determine a new White list by clustering.

Clustering process will be tested taking a universe of 550 users.

In order to perform these two runs we use WEKA and its module EXPLORER that offers several tools and models, as statistical as those used for classification and clustering [3]. The tool used to processes the data shows a graphical result in two dimensions: (customer_name, destination) so one cluster generated gives a particular White list, the idea is to get a minimum number of White lists, by reviewing the graphical dispersion of data in every cluster we see the convenience of either generate more clusters or keep only two clusters defining only two White lists. One more case is if we get one cluster with high concentration of destinations and a second cluster containing a few amount of destinations then we can decide keep one White list defined by the cluster with high concentration and the destinations conforming the second cluster would be kept out of the white list considering the amount of calls done to these destination during the six months period.

IV. Experimental Results

Performing the clustering process is simple through the interface of WEKA, we choose *SimpleKMeans*, the WEKA's k-means implementation as the preferred method and EM algorithm as a method to contrast against to. We choose the option "only without replacing missing values" in every run to generate clusters.

A. Experiment 1: the universe of users

The data sample to mine is composed by a universe of 550 users and 131 destinations. Some questions come up for this experiment of some practical interest:

Are there clusters with low traffic destinations that can be dropped? If so, then keep clusters with the strictly needed amount of destinations opened.

What is it worth to consider clusters to serve as inputs of a machine learning system to detect toll fraud?

B. Analysis and interpretation of data results

This section shows the results of Experiment 1, WEKA data mining tool gives the option to generate it graphically, so the dispersion of destinations is verified directly on the graph, this way is possible to check the behavior of destinations when the number of clusters grow, we want to know if the graphic dispersion of destinations grow then may be better to have more White lists, if not, then we need one or two clusters only, hence the same number of White lists.

In the following charts, the horizontal axis represents the number of clusters and the vertical axis represents the number of destinations, the vertical orientation shows the spatial distribution of users and destinations by cluster, they both conforms a cluster that takes the form of a column of dots.

K-means tests

First 3 runs for *SimpleKMeans*, are to generate 2 clusters, given 10, 11 and 12 seeds. Second 3 runs are to generate 3 clusters, given 10, 11, and 12 seeds.

Parameters 1: 2 clusters, 10 seeds

TABLE. 1. Results of experiment 1 with parameters 1

Cluster 0	**Cluster 1**
24 instances	107 instances
18%	82%
5 iterations, Square error = 2746	

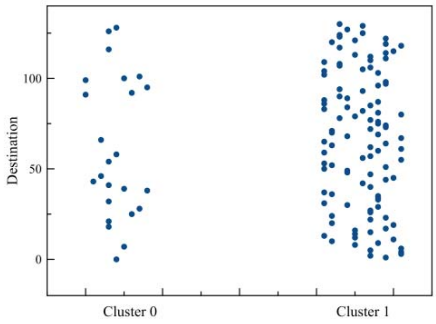


Fig. 5 K-means, 2 clusters, 10 seeds

From both the table and graph is evident that cluster 0 is the smaller one, destinations and users, the resulting White list for this cluster is

TABLE. 2. Resulting White list

Germany	Canada	Chile
Colombia	Costa Rica	Ecuador
USA	El Salvador	Spain
France	Guatemala	Honduras
Italy	Monaco	Mongolia
Puerto Rico	United Kingdom	Dominican Republic
Switzerland	Uruguay	Venezuela

All of the destinations, except Monaco and Mongolia, are well known to be over the mean of most called destinations, so this White list must be kept.

Parameters 2: 2 clusters, 11 seeds

TABLE. 1. Results of experiment 1 with parameters 2

Cluster 0	**Cluster 1**
112 instances	19 instances
85%	15%
3 iterations, Square error = 2755	

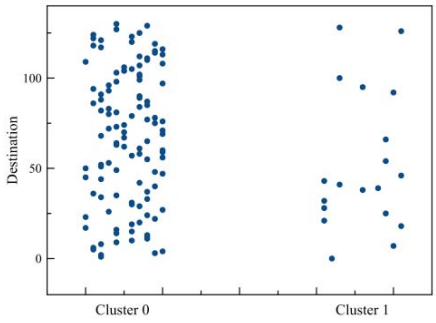


Fig. 6 K-means, 2 clusters, 11 seeds

Parameters 3: 2 clusters, 12 seeds

TABLE. 2. Results of experiment with parameters 4

Cluster 0	**Cluster 1**
24 instances	107 instances
18%	82%
3 iterations, Square error = 2755	

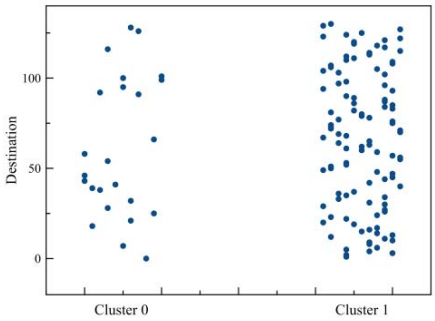


Fig. 7 K-means, 2 clusters, 11 seeds

Parameters 4: 3 clusters, 10 seeds

TABLE. 5. Results of experiment 2 with parameters 4

Cluster 0	**Cluster 1**	**Cluster 2**
83	29	19
63%	22%	15%
11 iterations, Square error = 2605		

Parameters 5: 3 clusters, 11 seeds

TABLE. 6. Results of experiment 2 with parameters 5

Cluster 0	**Cluster 1**	**Cluster 2**
106	20	5
81%	15%	4%
5 iterations, Square error = 2465		

Parameters 6: 3 clusters, 12 seeds

TABLE. 7. Results of experiment with parameters 6

Cluster 0	**Cluster 1**	**Cluster 2**
19	29	83
15%	22%	63%
11 iterations, Square error = 2605		

C. Experiment 2 EM algorithm

Following runs are intended to validate Experiment 1, As EM groups elements into every time bigger clusters, it is possible to run it without specifying number of clusters, instead a -1 is selected, to check the EM algorithm performance one additional run was done specifying 2 clusters, and a last run to compare against K-means.

Parameters 1: -1 clusters

TABLE. 8. Results of experiment 3 with parameters 1

Cluster 0	Cluster 1
28 instancias	103 instancias
21%	79%
Log likelihood: 357	

Parameters 2: 2 clusters

TABLE. 9. Results of experiment 3 with parameters 2

Cluster 0	Cluster 1
28 instances	103 instances
21%	79%
Log likelihood: 357	

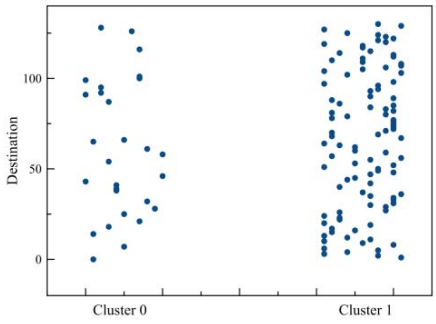


Fig. 8 EM algorithm, 2 clusters, log likelihood 357

Parameters 3: 2 clusters

TABLE. 3. Results of experiment 3 with parameters 3

Cluster 0	Cluster 1	Cluster 2
102 instances	21 instances	7 instances
79%	16%	5%
Log likelihood: 373		

It is observed that for 3 clusters, smaller clusters make no big difference so is not worth to have more White list. However the dispersion of destinations is different and it talks about different behavior, this way is interesting consider different amount of clusters to feed a machine learning to detect fraud, as some of these may reveal fraud traffic not easy to be seen by an operator during his duties.

From experiments 1 and 2 for K-means, we can see that the bigger cluster tends to be the same, even for 3 clusters the number of instances per cluster tend to keep the bigger cluster as the same. Experiment 2 for EM algorithm shows similar results validating the K-means method.

V. Conclusion

The analysis based on the clustering technique to define pattern behavior of traffic to international destinations from users VoIP has been useful to verify that a user does not need to have all of the worldwide destinations even though his business profile says that it needs to, so this kind of businesses can have more restrictive White lists. In regards to White list generation process we can state that having less White list is better than having lots of it, even individual White lists by user, going to the extreme. Clustering technique also allow us to create clusters hence White lists that define the behavior of a group of users, based in the history of changes in a six month period, that can be changed dynamically hence automatically.

More interesting is to see that clustering technique allow us to reveal similar behavior of users of long distance service, so it can be used to feed a machine learning system intended to detect fraud.

Acknowledgment

We would like to thank Nelva Nely Almanza Ortega and Antonio de Jesús Hernández Gómez (students of the Phd and MSc program at the Centro Nacional de Investigación y Desarrollo Tecnológico, CENIDET) for their assistance in the graphing of results and the literature search.

References

[1] Fawcett, T., and F. Provost. "Adaptive fraud detection. Data Mining and Knowledge Discovery", pp. 291–316. 1997

[2] Sengar, H., Wijesekera, D., Wang, H., Hahodia, S. "VoIP Intrusion Detection Through interacting Protocol State Machines", Center for Secure information Systems, George Mason University.

[3] Machine learning group at the University of Waikato. http://www.cs.waikato.ac.nz/ml/index.html

[4] Hollm´en, J., "User profiling and classification for fraud detection in mobile communications networks". ISBN 951-22-5239-2. (also published in print ISBN 951-666-555-1, ISSN 1456-9418). viii + 47 pp. UDC 004.032.26:519.21:621.391.

[5] R.Venkatesan, R. Ganesan, A. Arul Lawrence Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection" International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-7 December-2012, pp. 29-34

[6] Ogwueleka Francisca Nonyelum, "Fraud Detection in Mobile Communications Using Rule-Based and Neural Network System". Department of Mathematics, Statistics & Computer Science, University of Abuja, Abuja, FCT, Nigeria. 2010

[7] Grosser, H, Britos, P., Sicre, J., Servetto, A., García-Martínez, R., Perichinsky. G.4,3, "DETECCION DE FRAUDE EN TELEFONIA CELULAR USANDO REDES NEURONALES", CACIC 2003 – RedUNCI, [Online] Available: http://sedici.unlp.edu.ar/bitstream/handle/10915/22640/Documento_completo.pdf?sequence=1

[8] Moore, Andrew, "K-means and Hierarchical clustering", tutorial. [Online]. Available: http://www.autonlab.org/tutorials/kmeans.html

[9] Wiens, A., Wiens, T., Massoth, M., "A new Unsupervised User Profiling Approach for Detecting Toll Fraud in VoIP Networks", AICT2014: The Tenth Advanced International Conference on Telecommunications, pp. 63-69

[10] Kubler, S., Wiens, A., Wiens, T., Massoth, M., "Toll Fraud Detection in Voice over IP Networks using communications Behavior Patterns on Unlabeled Data", ICN2015: The Fourteenth International Conference on Networks, pp. 191-197.

[11] Abdallah A., Maarof, M., Zainal, A., "Fraud detection system: A survey", journal of Network and Computer Applications, 2016 Elsevier Ltd. pp. 91-104

[12] Sanver, M., and Karahoca, S., "Fraud Detection Using an Adaptive Neuro-Fuzzy Inference System in Mobile Telecommunication Networks", Journal of Mult.-Valued Logic & Soft Computing, Vol. 15, pp. 155–179

[13] Berkhin, P., "Survey of Clustering Data Mining Techniques", [Online] Available: http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf

[14] Cahill, M., Lambert, D., Pinheiro, J., Sun D., "Detecting fraud in the real world" , Handbook of massive datasets, 2002 Kluwer Academic Publishers. Volume 4 of the series Massive Computing. pp-911-929

[15] Becker, R., Volinsky, C., Wilks, A., "Fraud Detection in Telecommunications: History and Lessons Learned", Technometrics, Volume 52, Issue 1, 2010, pp. 20-33.

Modeling Maternal Mortality Rates in South Sudan

Gabriel Makuei, Mali Abdollahian, Kaye Marion
School of Mathematical and Geospatial Sciences
Royal Melbourne Institute of Technology (RMIT) University
Melbourne, Australia
G.makuei@student.rmit.edu.au, mali.abdollahian@rmit.edu.au, kaye.marion@rmit.edu.au

Abstract— **The maternal mortality rate (MMR) and the maternal deaths (MDs) in South Sudan is one of the highest in the world. The paper explores and compares the trends in HIV$^+$/AIDS and Non-HIV$^+$/AIDS related MMR. The results indicated that there is a declining trend in MMR. However, the decline in HIV$^+$/AIDS MMR is much slower than the decline in Non-HIV$^+$/AIDS MMR. The paper also aims for the first time to explore and compare the application of Log Linear and Poisson regression models to estimate MDs in South Sudan. Accuracy criteria such as coefficient of determination, Mean error are used to compare the predicting error of these models. The results show that Log Linear can model the MMR much better than Poisson.**

Keywords- maternal mortality , trend analysis, Log Regression, Poisson Regression, R^2.

I. INTRODUCTION

South Sudan attained its independence from Sudan in 2011. Although endowed with rich oil reserves, the country capped its oil wells recently due to disputes with Sudan. Thus, a major source of funds for its development has been blocked. Simultaneously, internal conflicts deter international investors from undertaking or supporting long term developmental projects in the country. One important sector affected by these problems is its health care sector, which essentially relates to human well-being and development. Maternal mortality rate (MMR) has been defined as the number of maternal deaths per 100,000 live births. This is the most commonly used measure for maternal mortality rate [1].

According to the data by World Health Organization (WHO) [1], although maternal mortality rate (MMR) in S. Sudan decreased from 1,000 per 100,000 live births to 730 per 100,000 live births during 2005 to 2013, it is still the highest in the world. The problem becomes more serious when considering the high fertility rate, weak health care system and high incidence of Human Immune Deficiency virus (HIV$^+$/AIDS) in the country.

Downie [2] gave the figure as 2,050 per 100,000 live births. Combined with high fertility rates, the probability of an average South Sudanese woman dying during one of her pregnancies is one in seven. South Sudan also has a significant female population with HIV$^+$/AIDS [4].

According to Calvert [5], HIV-infected pregnant women have eight times the risk of mortality than HIV-uninfected pregnant women. Based on this, about 25% of total MMR was attributable to HIV in Sub-Sahara. In Malawi and Zimbabwe, MMR increased by 1.5 and 2.5 times respectively along with a 10-fold increase of HIV incidence [6]. The authors point out that obstetric risk increases when quality of delivery of health services deteriorates. Even if safe motherhood programs exist, HIV-related illnesses may increase due to crowding in health facilities thereby affecting quality of health services. While this may be a factor of concern for South Sudan, also, there may be little change in the utilization of ante-natal care by pregnant women. Mugo, Dibley, & Ago [7] observed from South Sudan survey data for 2010 that only about 40% of the pregnant women used antenatal services and the frequency was less than the recommended four visits. Only about 18% of the pregnant women visited four or more times. Such a low level of antenatal services utilization is sure to affect efforts to reduce MMR. The need for research and evaluation to create evidence on HIV-related MMR in Sub-Saharan Africa was highlighted by Kendall, et al. [8]. Clinical management of pregnant and post-partum HIV affected women, effect of expanded antiretroviral therapy on maternal mortality and morbidity, integrated service delivery models and interventions to enable women with a social environment of continuous care are suggested as the areas for this study. In the case of South Sudan the results of this study show that the country lacks in all these respects The findings of Li, et al. [9] also higher MMR with HIV$^+$ and recommended initiation of antiretroviral therapy as early as possible during pregnancy.

Most of the above authors point out to the fact that HIV$^+$ is not the direct cause of MMR except in cases like sepsis. However, the association of higher MMR with HIV is a matter of concern. As there is no direct relationship, healthy women need not be prevented from becoming pregnant if they desire so. Rise of healthy deliveries will result in decrease of MMR. The importance of attending to socio-economic factors to reduce MMR was highlighted an Indian work by Rai & Tulchinsky [10].

The continued internal and external conflicts have destructed even the hospitals operation tasks [3]. Lack of humanitarian emergencies mostly in difficult-to-access areas, poor infrastructure, poor illiterate low-skilled population, low agricultural production, weak health care systems and shortage of technically capable medical staff are all factors and the causes of the largely unmet health care demands. High rates of both communicable and non-communicable diseases are reported even if there is no reliable health statistics data. Poor literacy levels act as barrier to improving health awareness to the tradition-bound population.

Lack of access to health care facilities is a major factor due to lack of roads and transport system [3]. South Sudan has

about 1,147 health care facilities that function to serve around nine million populations. Out of the facilities, only 37 are hospitals. As more than 50% of the population needs to walk three miles or more to the nearest primary health care unit, it not surprising that outpatient visit rate is only 20% per year. Ill-equipped buildings with poor hygiene are the common feature of these primary health care units. Chronic shortage of health care professional staff at all levels is demonstrated by 1.5 doctors and two nurses per 100,000 people. Health departments of the state and central governments are managed by poorly qualified personnel.

All the above factors affect the total health care system and in particular high maternal mortality rate problem.

One of the aims of this paper is to explore and compare the trend in the HIV$^+$/AIDS, non-HIV$^+$/AIDS MDs, and total maternal mortality rates between January 1986 and October 2015 through the data collected from one of the major health care referral center in South Sudan (Juba Teaching Hospital{JTH}, Juba, South Sudan). Juba Teaching Hospital is a 580-bed facility located in Juba City and is the biggest referral hospital in the whole country. The hospital is directly funded by the central government through the National Ministry of Health (NMoH), and supported by Risk Management Foundation (RMF), United Nation (UN) agencies and others (World Bank, United States Agency for International Development (USAID), and World Health Organization (WHO)). The results show that in general MMR is declining. However, the HIV$^+$/AIDS MMR is declining at a much slower rate compared to the Non-HIV$^+$/AIDS MMR.

Since the causes of death for HIV$^+$/AIDS and non-HIV$^+$/AIDS Maternal Deaths (TMDs) are different, the authors have decided to separate the HIV$^+$/AIDS and Non-HIV$^+$/AIDS total maternal mortality data for further statistical analysis. In the following sections we only report the analysis of the Non-HIV$^+$/AIDS Maternal Deaths data

This study for the first time aims to examine and investigate suitability of the Multi-log regression and Multi-Poisson regression models for estimating and detecting changes over time in rates of Non-HIV$^+$/AIDS maternal deaths in South Sudan. Based on the recommendation in the literatures listed above, the independent variables included in the analysis are Skilled Attendant at Birth (SAB), General Fertility Rate (GFR), and Gross Domestic Product (GDP). Accuracy criteria such as coefficient of determination, Mean error and standard error of mean are used to compare the predicting error of these models. The results show that Log Linear Regression can model the Maternal Mortality Death much better than Poisson Regression.

A reliable model to estimate the maternal deaths would assist the authorities to make an inform decision on resource allocation and lacking resources.

Methods

Time series plot and trend analysis are often used to observe patterns and structures in data over time. In this paper we have used Statistical package R to carry out trend analysis and model fitting to HIV$^+$/AIDS, non-HIV$^+$/AIDS, and total maternal mortality rates.

Regression modelling is a useful technique to model the strength and direction of relationship between one or more independent variables and a dependent variable. In this paper, multi- log regression and multi-Poisson regression have been utilized to gain insights into the predictors of non-HIV$^+$/AIDS Maternal Mortality Rate (MMR) (i.e. The independent variables that deployed into the model include Skilled Attendant at Birth (SAB), General Fertility Rate (GFR), Gross Domestic Product (GDP). The data used for this analysis are the aggregated data at the yearly level (1986 to 2015). (Source: Juba Teaching Hospital{JTH}, S. Sudan).

II-a Multi- Log Linear Regression model

In its general form, the linear regression model can be expressed as:

$Y=f(X_1, X_2, X_3, \ldots, X_{p)+}\, e$

Where Y is the response variable and $X_{1, \ldots, } X_p$ are p predictors, f is the function which links the predictors to the response, that its general form is a linear combination of predictors and e is error representing the discrepancy in the approximation (Montgomery et al., 2012). Using the yearly data from 1986-2015 for South Sudan , we have developed the following log linear regression model to describe the changes in Total Non-HIV$^+$/AIDS Maternal Deaths {TMDs} Rate in terms independent variables(IVs), SAB, GFR and GDP. Statistical package Minitab 17 and R are used to fit the best model.

The following Log Linear regression model was the outcome of the analysis based on 2/3 of the data:

Multi-Log Regression Equation, R^2 *=77.11%*

Log (Non-HIV/AIDS) = -20.8 - 8.30 Log (SAB) + 8.10 Log (GFR) + 5.12 Log (GDP).., (1)

We have modified the data provided by (WHO, UNICEF, UBFPA, the World Bank, and United Nations Population Division Maternal Mortality Estimation Inter-Agency Group in 1986-2015) to obtain a yearly value (rather than 5 years value) for independent variables (IVs) of SAB, GFR and GDP. The models are built based on randomly selected 2/3 of the data to overcome the decrease trend of the TMDs over the years. The remaining 1/3 will be used to assess the efficacy of the proposed models.

II-b Poisson Regression model

The Poisson regression model expresses the natural logarithm of the outcome or incident over a particular period of time as a linear function of a set of independent variables.

A measure of the goodness of fit for the Poisson regression model is acquired by using the deviance statistic of a partial model against a fuller model.

The Poisson log linear model with explanatory Y is written as

$\text{Log}(Y) = \dot{\alpha} + \beta x$

When there is a set of independent variables, then the model becomes

$Log(Y) = \dot{\alpha} + \beta X$

Where, the row vector β represents the coefficient factors and column vector X represents the independent variables (IVs). The following Poisson regression model was the outcome of the analysis based on 2/3 of the data:

Multi-Poisson Regression Equation, $R^2 = 79.75\%$

Non-HIV+/AIDS MDs Rate Per 1000 = exp(Y')

$Y' = 4.227 - 0.3819\ SAB + 0.03237\ GFR + 0.002902\ GDP, \ldots\ldots\ldots\ldots,(2)$

III. STATSICAL ANALYSIS

The statistical analysis in this paper includes trend analysis, time series modeling, multi-log linear regression and multi-Poisson regression.

III-a Time series analysis

Time series plot and summary statistics including means and standard deviations and box plots have been produced for yearly level of HIV$^+$/AIDS, non-HIV$^+$/AIDS and total MMR from 1986 to 2015 and are presented in Table 1 and Figures 1-2.

A linear trend model was fit to each individual time series of HIV$^+$/AIDS MMR, non-HIV$^+$/AIDS MMR, and the total MMR. The models are presented in Figures 3-5. The slopes of the models were compared to the slope of the total maternal mortality rate model. The fitted models are presented in table 2

The results in table 1 show that the mean HIV$^+$/AIDS MMR for the period of 1986 to 2015 were almost one third of the total MMR. The balance of the MMR was attributed to Non-HIV$^+$/AIDS related causes.

Table 1: Presents Summary Statics for HIV$^+$/AIDS, non-HIV$^+$/AIDS and total maternal mortality rate.

Variables	Mean	SD
HIV$^+$/AIDS MMR	1014.61	579.21
Non-HIV$^+$/AIDS MMR	2344.32	1466.39
Total MMR	3358.94	1674.88

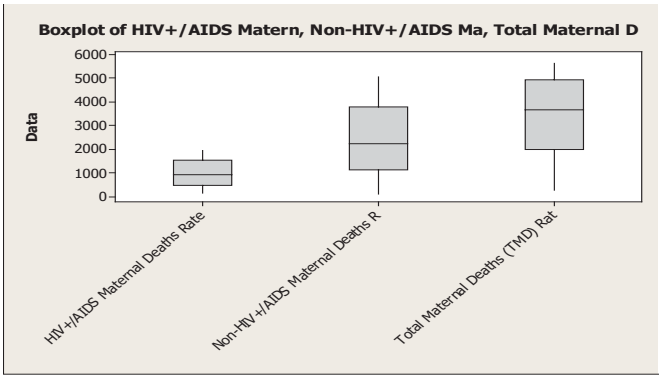


Figure 1: Box plots for yearly HIV$^+$/AIDS, non-HIV$^+$/AIDS and total MMR over the period of 1986-2015.

The three time series are shown in the figure below.

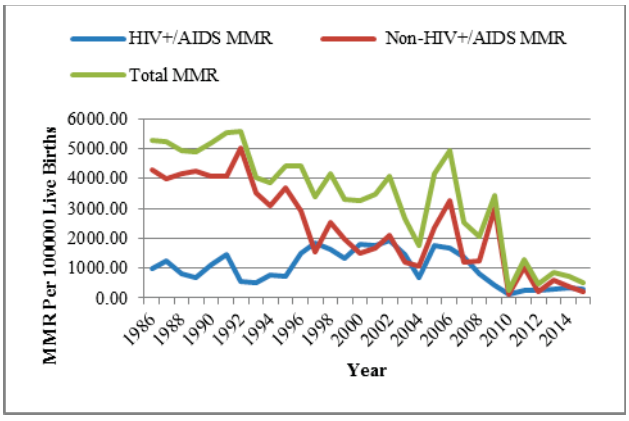


Figure 2: Trend Comparisons for yearly HIV$^+$/AIDS, non-HIV$^+$/AIDS and total maternal mortality over the period of 1986-2015.

Table 2: Time series models fitted to yearly HIV$^+$/AIDS, non-HIV$^+$/AIDS and total maternal mortality rate.

Time Series	Linear Trend Model
HIV+/AIDS MMR	Yt = 1337 - 20.8t
Non-HIV+/AIDS MMR	Yt = 4585 - 144.6t
Total MMR	Yt = 5992 - 165.3t

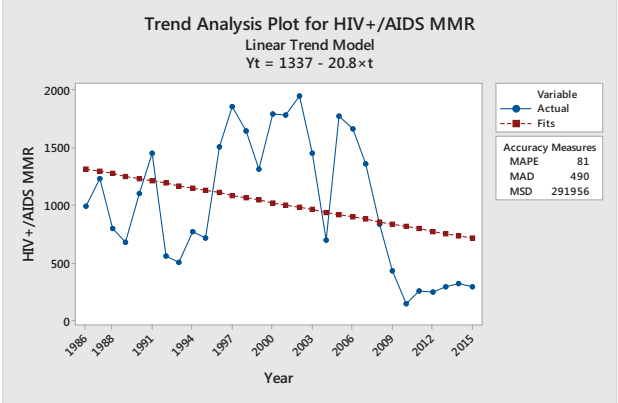


Figure 3: Trend analysis for yearly HIV$^+$/AIDS maternal mortality rate over the period of 1986-2015.

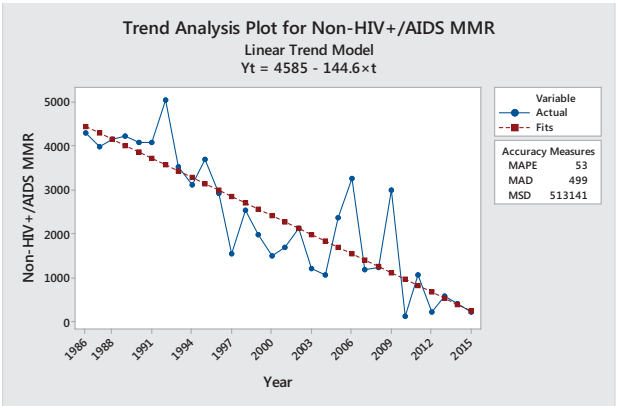


Figure 4: Trend analysis for yearly non-HIV$^+$/AIDS maternal mortality rate over the period of 1986-2015.

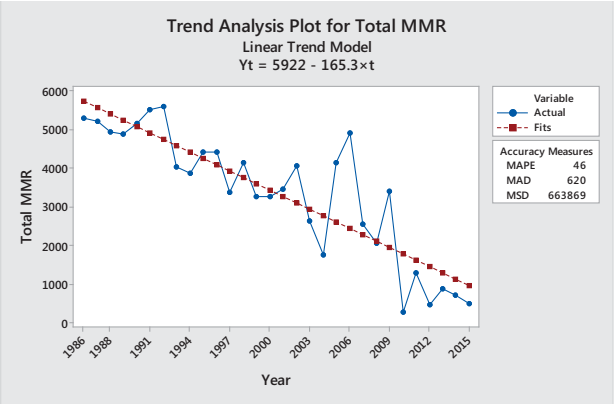


Figure 5: Trend analysis for total yearly maternal mortality rate over the period of 1986-2015.

Figures 3-5 show that the HIV$^+$/AIDS MMR linear trend model has a slope of 1,337, the Non-HIV$^+$/AIDS MMR linear trend model has a slope of 4,585, and the Total MMR linear trend model has a slope of 5,992. These results indicate that the difference between the slopes of the Non-HIV+/AIDS MMR and the Total MMR series (1,407) is much less than the difference between the slopes of the HIV+/AIDS MMR and the Total MMR series (4,655). Taking into consideration the declining trend in the three time series and the differences in the slopes of Non-HIV+/AIDS MMR and the Total MMR and HIV$^+$/AIDS MMR and the Total MMR, it would be safe to conclude that the HIV$^+$/AIDS MMR is declining at a much slower rate compared to the Non-HIV$^+$/AIDS MMR.

The summary statistics for the three time series by the year groupings are shown in the table .3 with the comparison graph in Figure 6.

Table 3: HIV$^+$/AIDS, Non- HIV$^+$/AIDS and Total MMR by Year Grouping

	Variable / Year	N	Mean	Min	Max	SD
HIV$^+$/AIDS MMR	1986 -2008	23	1236.46	504.41	1950.81	468.66
	2009 - 2015	7	285.67	145.31	434.45	87.02
	Total	30	1014.61	145.31	1950.81	579.21
Non-HIV$^+$/AIDS MMR	1986 - 2008	23	2816.27	1051.71	5042.02	1246.42
	2009 - 2015	7	793.64	121.09	2990.03	1018.60
	Total	30	2344.32	121.09	5042.02	1466.39
Total MMR	1986 - 2008	23	4052.73	1752.85	5602.24	1103.99
	2009 - 2015	7	1079.31	266.41	3424.48	1087.05
	Total	30	3358.94	266.41	5602.24	1674.88

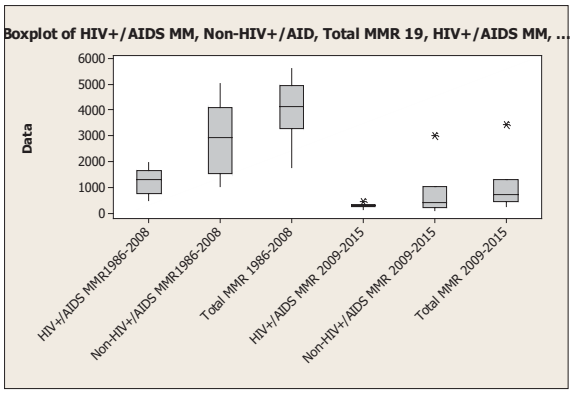


Figure 6: Comparison of HIV$^+$/AIDS, Non- HIV$^+$/AIDS and Total MMR by Year Grouping.

The results in the table .3 indicate that between the periods of (1986 -- 2008) and (2009 – 2015) the decline in HIV$^+$/AIDS maternal mortality rate has been higher (28.3%) compared with the decline in the Non-HIV$^+$/AIDS MMR (60.2%). This confirms the findings from the trend analysis which are presented in Figures 3, 4. However the numerical and graphical comparisons given in table .3 and Figure 6 show a significant decrease in mortality rate in the period of 2009-2015. Since the causes of death for HIV$^+$/AIDS and Non-Maternal Deaths (TMDs) are different the authors have decided to model them separately. Due to the limitation constrain, here we only outline the regression analysis of the Non-HIV$^+$/AIDS Maternal Deaths (TMDs).

III-b Regression Models for non-HIV$^+$/AIDS Maternal Deaths (TMDs)

As mentioned earlier, to establish Log Linear and Poisson Regression models, we used randomly selected two third of the Yearly Data to build the models. The models are then used to predict the remaining ten (10) years' data for Total Non-HIV$^+$/AIDS Maternal Death (TMDs). The analysis was carried out using Microsoft Excel, R and Minitab version .17 statistical soft wares.

The results of the prediction errors are presented in the table 4, and Figures 7-8. Table 4 indicates that the mean error percentage and the SE Mean for the Log linear regression is much smaller than Poisson regression. Therefore we can conclude that Log linear regression outperforms Poisson regression in predicting the morality data for South Sudan.

The results of error analyses form the following regression models:

Log Linear Regression Equation, R^2 =77.11%

Log (Non-HIV/AIDS) = -20.8 - 8.30 Log (SAB) + 8.10 Log (GFR) + 5.12 Log (GDP), ……..(1)

Poisson Regression Equation, R^2 = 79.75%

Non-HIV+/AIDS MDs Rate Per 1000 = exp(Y')

Y' = 4.227 - 0.3819 SAB + 0.03237 GFR + 0.002902 GDP,……(2)

Table 4: Presents errors analysis based on two third and one third of Yearly Data for independent variables(IVs) of: SAB, GFR and GDP.

Model	Mean Errors	SE Mean
Log Linear regression (1)	*0.008*	*0.171*
Poisson regression (2)	*-216*	*316*

The dependent variable: Non-HIV$^+$ /AIDs MDs (TMDs)

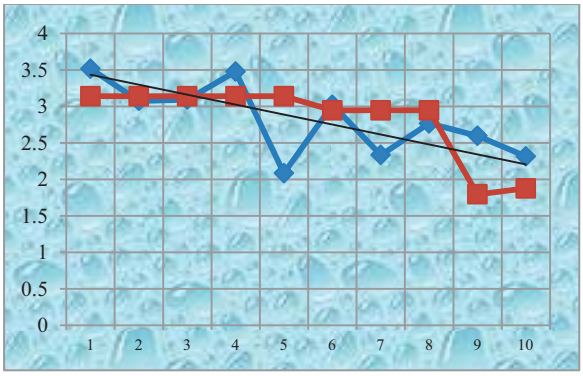


Figure 7: Presents actual and estimated values of one third of Non-HIV$^+$/AIDs MDs based on the independent variables SAB, GFR and GDP, using log linear regression.

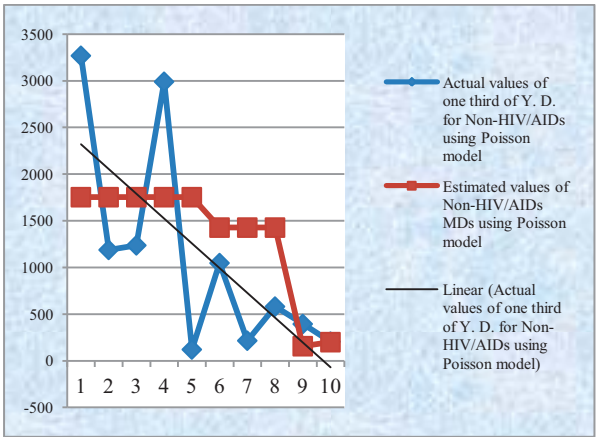


Figure 8: Presents actual and estimated values of one third of Non-HIV$^+$/AIDs MDs based on independent variables SAB, GFR and GDP, using Poisson regression.

IV. Discussion

The overall maternal mortality rate has been declining in South Sudan. However, the MMR in South Sudan is still one of the highest in the world ([1], [4]). Breaking down the MMR in the form of HIV$^+$/AIDS MDs based and Non-HIV$^+$/AIDS MMR and analysis of the trend serves many purposes, including: providing a magnitude of the HIV$^+$/AIDS MDs based MMR and providing insights into the trend in HIV$^+$/AIDS MDs based and Non-HIV$^+$/AIDS MMR. Such information is vital for the health policy makers in South Sudan and also for not-for-profit organizations like the United Nation organizations. The analysis will provide an effective decision making tool for formulating optimal strategies and resource allocation to address the issue of high MMR in South Sudan.

The result of the analysis have indicated that HIV+/AIDS MDs based MMR is a substantial contributor of the overall MMR in South Sudan with almost one-thirds of the maternal mortality deaths being attributed to HIV$^+$/AIDS MDs related causes. A further cause of alarm is that the HIV$^+$/AIDS related MMR is declining at a much slower rate compared to the overall MMR. This indicates that in the short to long term future, the MMR attributed to HIV$^+$/AIDS related causes might become a major contributor to the overall MMR.

There are many causes of the general high prevalence of HIV$^+$/AIDS in the country including low levels of education in large proportions of population. Frequent internal and external conflicts (e.g. wars) in the country are also contributors to the high prevalence of HIV$^+$/AIDS in the country. Internal and external conflicts plagued S. Sudan from 1983 to 2005, and again from 2013 to the present. This has caused a mass population displacement from within the country and from outside the country. Moreover, many people from neighboring countries like Uganda, Kenya, Ethiopia and Sudan have moved into South Sudan. As the incidence of HIV$^+$/AIDS in these countries are also high, therefore, it is safe to conclude that these people may have contributed to the high prevalence of HIV$^+$/AIDS based MMR in South Sudan.

The need to improve and increase evidence for effective interventions for reducing mortality among pregnant women with HIV$^+$/AIDS were stressed by Kendall, et al. [4]. However, to overcome this challenge we need better quality data on causes and factors of deaths among such women and enhanced and harmonized monitoring of the current health care programs. The authors have dedicated separate section of this project to this task and results will be reported in future papers. In the second part of the current research we have modeled the Non-HIV$^+$/AIDs maternal deaths based on the Yearly Data of independent variables SAB, GFR and GDP, using log linear regression and Poisson regression. The accuracy of prediction is pivotal to ensure that the estimate and forecast correctly reflects the future data. In this study, we have used coefficient of determination R^2, mean error and standard error of the mean to evaluate and assess the efficacy of the proposed models. The results of the analysis presented in table 4 together with the coefficient of determination R^2 show that log linear regression model based on independent variables (IVs) SAB, GFR and GDP can explain 77.1% of the variation in maternal deaths with the mean prediction error of 0.008.

For the future strategy management and action planning of maternal mortality rate (MMR) reduction in South Sudan, more accurate forecast models will be developed. One way to increase the prediction accuracy is to incorporate more independent variables in the models.

V. CONCLUSION

The need and urgency for the proposed work arises from the globally highest maternal mortality rate (MMR) reported for South Sudan. Although some efforts were made through policies and programs of the government and assistance by

various international agencies, reported reduction in mortality rate is only modest.

This study aims to compare the trends in HIV$^+$/AIDS MDs and Non-HIV$^+$/AIDS related mortality rates in South Sudan. Whilst it is observed that the HIV$^+$/AIDS MDs based MMR, Non-HIV$^+$/AIDS MMR, and overall MMR are on a general decline , the decline in HIV$^+$/AIDS based MMR is much slower compared to the overall MMR.

There is a possibility that in the near future the HIV$^+$/AIDS based MMR will be a bigger contributor to the overall MMR compared to Non-HIV$^+$/AIDS based MMR.

In recent years, there has been increasing interest in estimating and forecasting maternal deaths. Prior maternal death estimation can guide and assist the national and local governments to make inform maternal health care policies and medical resources allocation decisions especially in rural areas. Separating the Non-HIV$^+$/AIDs maternal deaths from the total maternal deaths, this paper has developed two models based on independent variables(IVs) SAB, GFR and GDP to estimate the maternal deaths. The models are developed using real data from the biggest referral hospital in South Sudan. The estimated maternal deaths were then compared with the recorded ones to evaluate the efficacy of these models. The accuracy criteria such as mean error and standard error of mean SE mean were used to compare the forecast errors of the models.

The analysis of the prediction error shows that the proposed multi-log linear regression model is cable of predicting the maternal death with minimum mean error and is out performing Multi-Poisson regression. The results also show that Skilled Attendant at Birth (SAB) is the most significant factor in decreasing the maternal death followed by based on independent variables**;** General Fertility Rate (GFR) and Gross Domestic Product (GDP).

The outcomes obtained from this study offer both challenges and opportunities for development of health care services as well as guide line for the resource allocation. Some of these may include improving the education levels and capacity building in medical fields in South Sudan.

Further research and evaluation are needed for improving clinical management of pregnant and postpartum pregnant women with HIV$^+$/AIDS and other Non-HIV$^+$/AIDS conditions.

The creation of a structure for an informative national data recording system is the most significant step in achieving the goal of this project. The authors had consultation discussions with 30 experts and review of literature to derive this conclusion.

ACKNOWLEDGMENT

This study was made as a part of doctoral degree by research. The authors would like to thank Dr. Maker Wal Lual, as a senior doctor and his devoted staff, for their effective effort in collecting the maternal deaths' data from the Biostatistics' Department, at the JTH, Juba, South Sudan.

References

[1] WHO, 2015. Countries: South Sudan. [Online] Available at: http://www.who.int/countries/ssd/en/ [Accessed 8 January 2015].

[2] Downie, R., 2012. *The state of public health in South Sudan*, s.l.: Centre for Strategic and International Studies, Global Health Policy Centre.

[3] Gorgeu, R., 2014. *South Sudan: Pervasive Violence Against Healthcare.* [Online] Available at: http://www.msf.org/article/south-sudan-pervasive-violence-against-healthcare [Accessed 8 January 2015].

[4] Kendall, T. et al., 2014. *Eliminating preventable HIV-related maternal mortality in Sub-Sharan Africa: What do we need to know?*. Journal of Acquired Immune Deficiency Syndromes, 67(Supplement 4), pp. S250-S258.

[5] Calvert, C. (2015). *The contribution of HIV to mortality in pregnant and postpartum women. London School of Hygiene and Tropical Medicine* (University of London).

[6] Bicego, G., Boerma, J. T., & Ronsmans, C. (2002). *The effect of AIDS on maternal mortality in Malawi and Zimbabwe*. Aids, 16(7), 1078-1081.

[7] Mugo, N. S., Dibley, M. J., & Agho, K. E. (2015). *Prevalence and risk factors for non-use of antenatal care visits: analysis of the 2010 South Sudan household survey.* BMC Pregnancy Childbirth, 15(1), 68.

[8] Kendall, T., Danel, I., Cooper, D., Dilmitis, S., Kaida, A., Kourtis, A. P., et al. (2014). *Eliminating Preventable HIV-Related Maternal Mortality in Sub-Saharan Africa: What Do We Need to Know?* Journal of acquired immune deficiency syndromes, 67(Supplement 4), S250.

[9] Li, N. M., Spiegelman, D., Chalamilla, G., Hertzmank, E., Sando, D., Sando, M. M., et al. (2014). *Maternal mortality among HIV‐infected pregnant women in Tanzania*. Acta obstetricia et gynecologica Scandinavica, 93(5), 463-468.

[10] Rai, R. K., & Tulchinsky, T. H. (2015). *Addressing the sluggish progress in reducing maternal mortality in India*. Asia-Pacific Journal of Public Health, 27(2), NP1161-NP1169.

Sentiment Analisis on Web-based Reviews using Data Mining and Support Vector Machine

Renato S. C. da Rocha
Department of Informatics
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil
rrsayao@gmail.com

Marco Aurelio Pacheco
Electrical Engineering Department
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil
marco@ele.puc-rio.br

Leonardo A. Forero Mendonza
Electrical Engineering Department
State University of Rio de Janeiro
Rio de Janeiro, Brazil
leofome@hotmail.com,

***Abstract*— This work aims to use sentiment analysis techniques, data mining, text mining and natural language processing to indicate the polarity of texts using support vector machine. Weka software and a movie review database from Internet Movie Database - IMDb - were used. This work uses preprocessing filters and WRAPPER techniques and Support Vector Machine (SVM) for classification. It presents better results when compared to other preprocessing techniques used in sentiment analysis.**

Keywords—text mining; sentiment analysis; machine learning; support vector machine; preprocessing techniques; filter; wrapper;

I. INTRODUCTION

Currently, the web user's behavior is changing. As well as consuming content available on the web, people are also exposing their opinions and experiences about products acquired, visited places and services they used. These reports may influence the decision of other users, serving as additional information that is not often available in the description of the product. This feedback process can be very useful both for companies, which may use this information to improve their products, and for consumers, who can also take advantage of the experience of other users.

Sentiment analysis area is dedicated to the analysis of opinions, feelings, emotions and attitudes of people and has become an object of research in various scenarios from social media such as blogs, e-mail, discussion forums, products review, etc. Opinion Mining is a part of text mining with focus on processing user-generated content. This feature adds several research challenges such as identifying topics and opinions. User-generated content are considered unstructured data, as they may deal with various subjects in the form of free text. In order to make the data more easily understandable, developed methods seek to process the opinions so that they can be represented in a structured way. Structured data allow a straightforward analysis of the main topics along with the given average opinion. With the growing popularity of social media, opinions processing tools have to deal with large amounts of data. Thus, it becomes imperative to represent data in a summarized form. The task of sentiment analysis is quite complex, firstly by language problems inherent in the problem, such as sentences formation and spelling rules. Secondly, depending on the application, for example, how to classify movie reviews, there are people who use irony and sarcasm, which should be identified by the algorithms. Thirdly is the data preprocessing and structuring. In this regard, classical techniques are applied, but rely heavily on the specialist's knowledge.

This paper presented a classification of a movie review database, using support vector machines as the classifier algorithm. For data preprocessing, classical techniques presented in various works were compared with a mix of classical filters and the wrapper preprocessing technique, which performed better in addition to avoiding specialist interference.

II. FUNDAMENTALS OF TEXT MINING

Text mining is an important step of Knowledge Discovery from Text (KDT) [1]. According to [2] text mining is a variation on data mining: while regular data mining extracts patterns, regularities or other trends from structured databases of facts, text mining leads with problem of natural language processing.

Gleaning useful information from natural language text, however, has been a daunting task because text is amorphous and unstructured. Unstructured information, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data [3].

Most unstructured data are not modeled, are random, and are difficult to analyze. In particular, text is far more complex, involving cultural nuances in the communication of information, opinions, or dramatic narrative. Nonetheless, with the advancements in data storage and the ready availability of digitized texts, text mining can help in acquiring better competitive intelligence.

Text mining uses techniques from information retrieval, natural language processing (NLP), statistics, machine learning, and specially data mining. The choice of the techniques depends on the dataset and the nature of text

This work has been financially supported by CNPq and FAPERJ.

mining task. Typical text mining tasks include text classification, text clustering, information extraction, and sentimental analysis [4].

Sentiment analysis (also called opinion mining, review mining, appraisal extraction or attitude analysis) is the task of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in textual input [5], [6].

The next subsections provides a brief introduction to opinion mining, as well as to the support vector machine (SVM) classifier that constitute the essence of this work.

A. Sentiment analysis

The sentiment analysis is the computational study of expressed opinions, often subjective, in any snippet of text in natural language (document). The entire opinion is composed of at least two basic elements: a target and a sentiment about the target [7]. A target consists of an object, also called entity or topic, or characteristics of the object, also called aspects, which can be a product, person, organization, brand, event, among others. The terms sentiment and opinion are often used synonymously in this context. The polarity of a sentiment may be classified into discrete classes (positive, negative or neutral) [8], or as a range that represents the intensity of the sentiment, typically $[-1,1]$. Besides entity, aspect and polarity, the other two characteristics that complete an opinion quintuple [7] are the opinion source and the time the opinion was expressed.

Thus, the main sentiment analysis task can be defined as follows: given a document D, identify the expressed opinions about an entity, its aspects and polarity. A document can be analyzed at different levels of granularity: i) the lower the granularity, the more specific the classification; ii) the decision level is subjected to the context and application. In this sense, an opinion can be classified as to its polarity in terms of: document, sentence, entities and aspects.

Opinions may be regular or comparative; direct or indirect, and implicit or explicit [7]. The way opinions are expressed directly influences the ability to properly process them. Because they are easier to be treated, most works focuses on regular, direct and explicit opinions. The challenges in the opinions processing, inherent in natural language processing, are: words disambiguation; sarcasm and irony; semantics and syntax, among others.

Some steps are necessary to perform the sentiment analysis, since text mining originate from multiple sources in various formats. In general a process of text mining occurs in five macro steps [9]: data collecting, preprocessing , indexing, mining and analysis.

The first step aims at gathering information to compose the textual database to work (corpus), i.e., it involves determining and selecting the universe on which the text mining techniques will be applied. Collecting social data is usually done with APIs (Application Programming Interface).

After collecting documents, the preprocessing step structures them for the automatic knowledge extraction algorithms application. Primordial to the entire mining process performance, the preprocessing operations include: sectioning of a document in minimum units with the original text semantics (tokens) and removal of tokens without semantic and irrelevant value for mining (stop words) [10].

There is also the possibility of applying statistical information based filters that influence the classification quality: IDFTransform method (Inverse Document Frequency Transform) takes the premise that the attributes that rarely appear are valuable for classification, while the TFTransform method (Term Frequency Transform) admits that the most common terms are more important. Additionally, the following NLP techniques are often used, improving results: order and position of words identification, grammatical classes labeling, speech analysis, reduction of derived words to their root form (stemming), and the conversion/correction of informal writing, abbreviations and emphasis on words by repeating characters, quite common in social networks and that produce inaccurate evaluations by traditional mining techniques.

Thereafter starts the indexing phase. Indexing is the process responsible for creating auxiliary structures called indexes that guarantee speed and agility in the recovery of documents and its terms. Two more efficient distinct approaches are present in the text mining works: textual indexing and thematic indexing [11].

Once indexed, documents and terms are subjected to machine learning algorithms to perform knowledge extraction (mining step).

Finished the mining step, the sentiment analysis of extracted messages is carried out. The goal is the positive, negative or neutral polarity classification.

B. Polarity Classification

Literature divides different sentiment classification techniques on three approaches: lexicon based approach, machine learning approach and hybrid approach [12].

The lexicon approach is based on a collection of sentiment terms previously known and pre-compiled and can be of two types: dictionary-based [13] and corpus-based [7]. A dictionary consists of a database comprising words previously classified according to their polarities, and can be constructed either manually or from other words, called seed words. In the corpus-based approach, the semantics technique is very similar to the statistical, except the polarity is measured in terms of some measure of distance between terms, often Pointwise Mutual Information (PMI) [14]. The techniques principle in this category is that semantically close words must have the same polarity.

In machine learning-based approach supervised methods are employed, classification to be more especific. Basically, these methods consist in the execution of two processes: i) learn a classification model on a training corpus with previously labeled classes (positive and negative, for example); ii) use the model obtained in i) to classify documents that were not used in the construction of the classifier. Support Vector Machine (SVM) is among the most successful algorithms in classification tasks [15].

SVM algorithm represents documents as points in a vector space, which dimensions are selected features. Using the document training vectors, the basic idea of SVM is to find the optimal hyperplane that separates the previously classified data with the largest margin of separation between the two classes. The optimal hyperplane is then used to classify unlabelled data. The support vectors are those that define the optimal hyperplane separation location. SVMs deal, very effectively with non-linear problems, mapping the training set of its original space to a new larger space, outperforming other techniques such as artificial neural networks [16]. Literature describes a wide range of SVM application in text mining tasks [6].

However, in high-dimensional feature space, supervised methods, such as SVMs, suffer due to the curse of dimensionality [17]. A possible solution to avoid this issue is to use feature selection techniques. They are often used to reduce the dimensionality of the feature space and improve computational efficiency and accuracy of classifiers [18]. One successful approach for feature selection, which fits very well the Web data extraction problem [19], is based on wrappers [20]. Wrappers search for an optimal feature subset using the classification accuracy of some learning algorithm as their evaluation function. Thus, the best search-fit is an optimization problem and, therefore, several techniques can be used, including the evolutionary algorithms. Evolutionary algorithms, such as genetic algorithms [21], are population-based metaheuristics of great research interest because of their promising results in different application. These metaheuristics use principles of Darwin's theory of natural selection: at each generation or iteration of the algorithm, a competitive selection occurs to choose the best solutions; these are modified by crossover and mutation operators to generate new solutions, repeating this cycle until a given stop criterion, defined by the user, is reached.

The next section describes the proposed method for classifiers construction.

III. Methodology

The sentiment classification task will be divided into three steps: information collecting is the first step, database preprocessing, the second step, and the third and last step is the database classification to find polarity of each test observation. Weka software was used to develop this work.

1- Database

An existing database was used, and the information collecting has been previously made. The database named 'newsgroup rec.arts.movies.reviews' from IMDb Internet Movie Database [5] contains several movie reviews, which were collected, classified between positive and negative and made available to test sentiment analysis algorithms. This database is quite complex from the sense of sentiment analysis (SA) since movie reviews contain ironies and sarcasms which can affect the performance of classification algorithms. The database consists of two thousand files divided into two groups: 1000 observations with positive polarity and 1000 observations with negative polarity.

Being a known and widely used benchmark, this database facilitates results comparison with other algorithms.

2- Data preprocessing

Firstly, the database was divided by tokens. Those tokens will be the text's words. Then, three different preprocessing configurations were used to compare the classification result and determine which technique is more efficient.

a) The first preprocessing methodology used the following techniques:

- IDFTransform
- LowerCaseTokens
- MinTermFreq
- StopWords
- Stemmer

Fig. 1 shows the steps to the first preprocessing methodology.

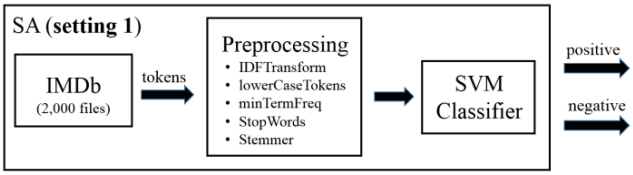


Fig. 1. First preprocessing methodology

b) A genetic algorithm wrapper was used in the second preprocessing methodology. The SVM classifier was the evaluation function and the classification accuracy was used to evaluate the generated solutions.

The genetic algorithm (GA) used the database set of words as the basic chromosome; each chromosome gene comprises a database word. Genetic mutation and crossover operators were used, with fixed rates of 0.3 and 0.6, respectively.

Fig. 2 shows the steps to the second preprocessing methodology.

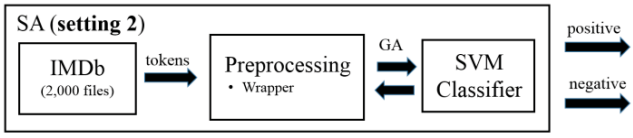


Fig. 2. Second preprocessing methodology

c) In the third preprocessing methodology, both filters presented on the first and wrapper presented on the second methodology were used.

Fig. 3 shows the steps to the third preprocessing methodology.

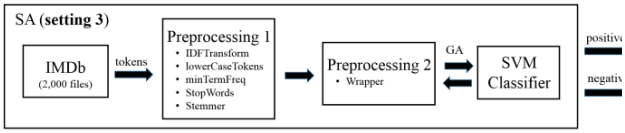


Fig. 3. Third preprocessing methodology

3- Database Classification

For each preprocessing methodology, a support vector machine classifier with two different kernels was used: polynomial and radial basis function (RBF). Experiments were made with different settings until the best configuration was reached. It was used 80% of the database for training and 20% for testing.

IV. Results

Section 3.2.a preprocessing methodology was used in the first experiment. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of σ and the complex variable are changed in the RBF kernel. Results are presented below in Table 1 and Table 2:

Table 1. RBF Kernel models for first preprocessing methodology

RBF Kernel				
	g=0.1	g=0.05	g=0.01	g=0.005
c = 1.0	48.6%	71.3%	87.6%	87.3%
c = 1.5	50.6%	73.3%	86.6%	87.6%
c = 2.0	50.3%	73.3%	86.0%	86.0%

Table 2. Polynomial Kernel models for first preprocessing methodology

Polynomial Kernel				
	e = 1.0	e = 0.5	e= 0.25	e= 0.1
c = 1.0	78.6%	82.0%	86.0%	88.6%
c = 1.5	77.6%	83.0%	85.0%	89.0%
c = 2.0	77.6%	81.0%	83.0%	89.3%

The best classification configuration with the first preprocessing configuration was obtained with exponent 0.1 and C = 2.0 with polynomial kernel, resulting in 89.3% accuracy.

The second experiment, section 3.2.b preprocessing methodology was used. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of σ and the complex variable are changed in the RBF kernel. Results are presented below in Table 3 and Table 4:

Table 3. RBF Kernel models for second preprocessing methodology

RBF Kernel				
	g=0.1	g=0.05	g=0.01	g=0.005
c = 1.0	49.4%	74.3%	88.6%	89.3%
c = 1.5	57.1%	75.3%	89.6%	89.6%
c = 2.0	57.2%	75.3%	86.0%	90.2%

Table 4. Polynomial Kernel models for second preprocessing methodology

Polynomial Kernel				
	e = 1.0	e = 0.5	e= 0.25	e= 0.1
c = 1.0	80.8%	84.0%	88.0%	88.9%
c = 1.5	77.6%	85.2%	91.1%	88.9%
c = 2.0	77.6%	85.2%	91.1%	88.9%

The best classification configuration with the second preprocessing configuration was obtained with exponent 0.25 and C = 1.5 with polynomial kernel, resulting in 91.1% accuracy.

The third experiment, section 3.2.c preprocessing methodology was used. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of σ and the complex variable are changed in the RBF kernel. Results are presented below in Table 5 and Table 6:

Table 5. RBF Kernel models for third preprocessing methodology

RBF Kernel				
	g=0.1	g=0.05	g=0.01	g=0.005
c = 1.0	61.4%	77.3%	90.4%	90.4%
c = 1.5	62.8%	78.9%	90.4%	91.6%
c = 2.0	62.8%	78.9%	89.8%	91.6%

Table 6. Polynomial Kernel models for third preprocessing methodology

Polynomial Kernel				
	e = 1.0	e = 0.5	e= 0.25	e= 0.1
c = 1.0	88.3%	88.2%	90.1%	93.7%
c = 1.5	87.9%	88.7%	92.6%	93,7%
c = 2.0	87.6%	88.9%	92.6%	92.6%

The best classification configuration with the second preprocessing configuration was obtained with exponent 0.1 and C = 1.5 with polynomial kernel, resulting in 93.7% accuracy.

V. Results Analysis

The third preprocessing configuration, which blends classical techniques of text mining approach to the wrapper, outperformed the other two models. This hybrid data preprocessing methodology was very efficient, and had better accuracy in almost every SVM configurations when compared with the other preprocessing methodologies.

The method shown in the second part of the results section shows how wrapper technique alone achieves better accuracy than the classic filter methods of text mining shown in the first part of results.

These results were compared with two other sentiment analysis classification works: [22] achieved 90.3% and [23] achieved 81%, both working with deep learning techniques, using the same database. The best result achieved in this work 93.7% performed better than this other two works.

By using the genetic algorithm wrapper methodology, this model has a slightly higher computational cost when compared to deep learning techniques, but the results are better and consistent. As it is an optimized model, it can be more robust to changes in the database than traditional methods.

VI. Conclusion

This paper has presented a sentimental analysis model, combining the wrapper method with the SVM classifier. This model has improved the text classification compared to other models using the same database as test.

The results show the wrapper-preprocessing filter can effectively clean the data. And when it is used jointly with classical preprocessing filters, it provides superior results. This technique is not found in sentimental analysis tasks, but it is often used in data mining. The classification task heavily depends on data cleaning. As stop words and stemmer are very subjective, the wrapper suffers less influence from specialist, being more robust.

References

[1] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," *LDV Forum - Gld. J. Comput. Linguist. Lang. Technol.*, vol. 20, pp. 19–62, 2005.

[2] M. A. Hearst, "What Is Text Mining?," 2003. [Online]. Available: http://people.ischool.berkeley.edu/~hearst/text-mining.html. [Accessed: 12-May-2016].

[3] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. Mahmoud Ali, M. Alam, M. Shiraz, and A. Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, pp. 1–18, 2014.

[4] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006.

[5] A. Montoyo, P. Martínez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments," *Decis. Support Syst.*, vol. 53, no. 4, pp. 675–679, Nov. 2012.

[6] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.

[7] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[8] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 478–514, May 2012.

[9] C. N. Aranha, "An automatic preprocessing for text mining in portuguese: a computer-aided approach," Pontifical University Catholic of Rio de Janeiro, Brazil, 2007.

[10] M. Konchady, *Text Mining Application Programming*, 1st ed. Rockland, MA, USA: Charles River Media, Inc., 2006.

[11] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books)*. Addison-Wesley Professional, 2011.

[12] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[14] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," in *Proceedings of the 27th annual meeting on Association for Computational Linguistics* -, 1989, pp. 76–83.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY: Springer New York, 2000.

[16] L. A. F. Mendoza, "Redes Neurais e Máquinas de Vetores de Suporte no Reconhecimento de Locutor usando Coeficientes MFC e Características do Sinal Glotal," Universidade Federal Fluminense, Rio de Janeiro, Brazil (in Portuguese), 2009.

[17] R. Bellman, *Dynamic Programming*, 1st ed. Princeton, NJ, USA: Princeton University Press, 1957.

[18] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009.

[19] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Syst.*, vol. 70, pp. 301–323, Nov. 2014.

[20] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in *Machine Learning Proceedings 1994*, Elsevier, 1994, pp. 121–129.

[21] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, vol. Addison-We. 1989.

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011, pp. 142–150.

[23] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," in *Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media*, 2009.

ISBN: 1-60132-441-3, CSREA Press ©

SESSION

POSTER PAPERS

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

Multi-dimensional Text Warehousing for Text Analytics

Jiyun Kim[1], **Han-joon Kim**[1]
[1]School of Electrical and Computer Engineering, University of Seoul, Seoul, Korea

Abstract - *A data warehouse is a repository of integrated data that provides the basis for decision making which is required to establish crucial business strategies. Recent data warehouses focus upon not only structured data but also textual data. Specially, a text warehouse is a type of data warehouse that provides efficient document retrieval and summarization capabilities. In this paper, we propose a novel way of text warehousing to support text mining capabilities beyond the simple search and aggregation functions. The proposed text warehousing method allows us to perform text mining tasks such as text classification, text clustering, and word association mining more effectively.*

Keywords: Data warehouse, Text warehouse, Text mining, OLAP

1 Introduction

Nowadays, a data warehouse that provides multi-dimensional OLAP capabilities is an important information repository for business intelligence. It can provide significant multi-dimensional views of target measures by integrating huge amounts of structured data retrieved from various sources of relational databases. Currently, due to explosive growth of textual data, many studies on text (or document) warehouses have been carried out to develop efficient document retrieval and summarization capabilities.

In this paper, we introduce a new way of building a text warehouse that provides text mining capabilities as well as simple search and aggregation functions. Our proposed method makes it possible to perform various text mining tasks such as text classification, text clustering, and word association mining. In order to support more reliable text mining tasks, we intend to extract concepts hidden in documents or words. For this, we employ the Wikipedia articles. The topological relationships among concepts are determined in our previous work using the search engine named *Elasticsearch*. Significant probabilistic weights for text mining are stored in the text warehouse, and they are used in conducting text mining algorithms such as naïve Bayes classification, EM clustering, and A priori-based word associations.

2 Backgrounds

Wikipedia is a free-content Internet encyclopedia, supported by the non-profit Wikimedia Foundation [1]. Each of Wikipedia articles allows defining a concept [2]. A Wikipedia article includes a title and a body text, and the body text contains an info-box and an anchor texts. These components are used in isolating good quality concept-level articles. In [3], G. Lee proposed a method of building corpus-dependent topic graphs using Wikipedia-based *Elasticsearch* search engine. With a relational database containing well-chosen Wikipedia articles, we can develop the concept network that depends on a given document corpus, and the concept network is materialized as a dimension table in our text warehouse.

3 Text Warehousing for Text Mining

3.1 Text warehouses

Our goal is to develop a text warehouse (TW) that provides text mining capabilities as well as simple search and aggregation functions. Figure 1 shows the proposed architecture for the proposed text warehouse.

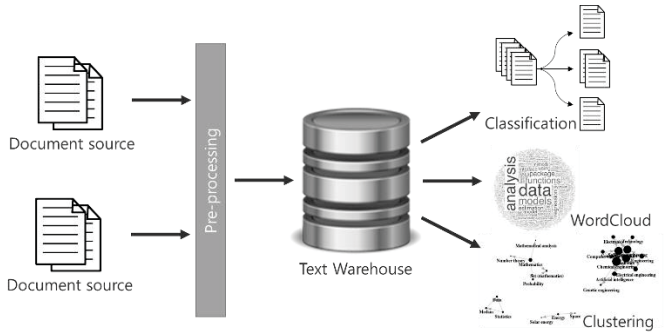


Figure 1. System architecture of the proposed text warehouse

Similarly to a general data warehousing, we firstly perform integrating document data from diverse data sources, and conduct pre-processing tasks such as keyword extraction, tokenizing, semantic tagging and document summarization. At this time, we record the metadata such as title, author, subject, date, and format. Our text warehouse is developed according to the multi-dimensional schema given in Figure 2. The important thing is that the sense of words occurring in incoming documents is saved with its corresponding weights in TW, and also the information of relevant words are saved in a dimension table of TW by using WordNet; WordNet is a lexical database for the English language, which consists of synonyms, short definitions and usage examples [4]. The developed text warehouse will be used for OLAP operations (such as drill-down and roll-up), and text mining tasks (such as text classification and clustering).

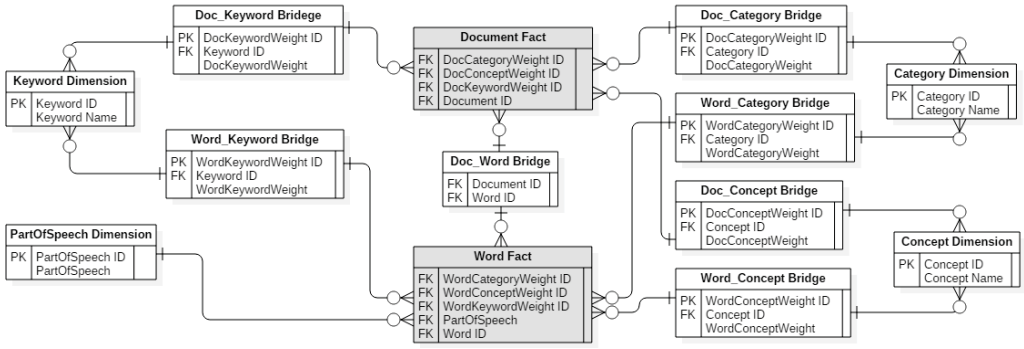


Figure 2. The proposed schema of text warhouse

3.2 Multi-dimensional modeling

Figure 2 shows the database schema of the TW that supports text mining tasks; the schema consists of two fact tables, several dimension tables and bridge tables. The 'document' fact table stores the documents collected and their metadata. The 'word' fact table stores the word information of the document collected. The 'document word' bridge table is located between 'document' fact and 'word' fact. These tables have a many-to-many relationship with each other, and the bridge table yields a one-to-many relationship with two facts [5].

The 'category' dimension table expresses the category of documents (or words). The 'part of speech' dimension table contains the information what type of word it is. The concepts extracted from the Wikipedia database are stored in the 'concept' dimension table, which has topological relationships among concepts. The keywords extracted from the WordNet database are stored in the 'keyword' dimension table, which has the metadata about keywords and their synonyms. Because each of dimension tables and the fact table are the many-to-many relationships, the bridge table is required between the tables.

3.3 Weight factors for text mining

As stated before, our proposed TW supports text mining tasks, and for this various types of weights are stored in the bridge table. In our work, the weights are computed in a probabilistic manner. As for text classification, Bayes' theorem can be used to find the maximum posterior probability, and then the probability that the document *d* is classified into the category *c* is simplified as follows:

$$P(c|d) \approx \prod_{w \in W} P(w_i|c)P(c) \qquad (1)$$

where w_i denotes each of words occurring in the document d.

For more accurate text mining tasks, we consider the sense (or meaning) of words. Thus Equation 1 is expanded as Equation 2 with the additional random variable; i.e., the concept (or sense) s_j that the word w_i has.

$$P(c|d) \approx \prod_{w \in W} \sum_{V_s} P(w_i|s_j, c)\, P(s_j|c)P(c) \qquad (2)$$

where s_j denotes a concept extracted from Wikipedia.
$P(w_i|s_j, c)$ corresponds to the 'WordConceptWeight' field of 'Word_Concept' bridge table, $P(s_j|c)$ corresponds to the 'DocConceptWeight' field of 'Doc_Concept' bridge table, and $P(c)$ corresponds to the 'DocCategoryWeight' field of the 'Doc_Category' bridge table. These weights significantly contribute to enhance text classification algorithms (such as naïve Bayes) and text clustering algorithms (such as EM).

4 Summary

In this paper, we have suggested a new way of building the text warehouses to support text mining tasks. Our proposed method emphasizes the sense of words occurring in documents, and thus all of related probabilistic weights are stored in the bridge tables. Here, the word sense is accounted for in the concept space defined with the Wikipedia. Therefore, we expect that our text warehouse makes it possible to perform most of text mining tasks more efficiently.

5 Acknowledgments

This research was supported by a grant (15AUDP-B100356-01) from Urban Architecture Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government, and was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF-2015R1D1A1A09061299) funded by the Ministry of Education.

6 References

[1] Wikipedia, https://en.wikipedia.org/wiki/Wikipedia.

[2] G. H. Lee and H. J. Kim. "Automated Development of Concept Hierarchy Tree using Backlink Information of Wikipedia"; KIISE SIGDB, Vol.31 No.1, 40-49, Apr 2015.

[3] G. H. Lee. "Incremental Development of Corpus-Dependent Concept Hierarchy Graph using Wikipedia", Master's Dissertation. University of Seoul, Korea, Feb 2016.

[4] WordNet, https://en.wikipedia.org/wiki/WordNet.

[5] R. Kimball and M. Ross. "The data warehouse toolkit: the complete guide to dimensional modeling". John Wiley & Sons, 2011.

Expanding Educational Horizon to Accommodate All Individuals through Lens of Deep Data Analytics

(Work in progress)

Muhammad Fahim Uddin
Graduate Student, School of Computer Science and Engineering, University of Bridgeport, CT. USA

Extended Abstract - *We have abundance of schools, colleges and universities around the world. Such institutions are backbone of any developed society and are core to growth and civilization. We provide almost every kind of educational areas in science and arts, broadly speaking. Basic schooling is everybody's right in any society regardless of their poverty or skill level. However, statistics show that not every individual will go through same level of education as of others. In addition to this, some of individuals don't get to area of education; their inner talent is made for. For example, it is not uncommon for a very technical person (by birth) ends up in non-technical education path and real world jobs. Such scenarios and examples are everywhere. This results in an inefficient distribution of education and skills to right individuals. To solve this problem, I pursue researching big and unstructured data from all sources that is related to education, career and talent. In such exploration, I also focus on social networking data and mine it to analyze hidden personality features that can contribute to understand different personalities. In this research, I explore various data and analysis techniques to implement algorithms and models through lens of cognitive computing and artificial intelligence. I aim to use a very huge data set, so machine learning and training data techniques can be implemented to correlate features. I believe such analysis and mining of data, identify new educational areas, curriculum and sectors, that we must introduce to ours schools and colleges, in order to provide very customized education to a very special individuals that normally don't fit in standard educational system and fail to retain normal journey. This research is in conjunction (and part of, result of) with our other research work/initiatives in data mining, personality prediction, educational data mining and artificial intelligence, which we are pursuing and sharing with community in journals and conference at present.*

1 Data Sources

I plan to get available data from Facebook, Twitter, LinkedIn and other sources including colleges and universities. I will use data collected from past students on campus and random interviews across campus.

2 Methodologies

I aim to enhance existing algorithms, combine them and create new algorithms to show improved data mining, data discovery and data analysis for the objective, discussed in abstract. I use Python and R for our data analysis. We use Microsoft SQL Server and Excel for data repositories and mining process.

3 Related Study

I am studying deep literature and publications of last 2 decades to understand the history and latest state of the art. I will provide survey of related tools and algorithms.

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

SESSION

LATE BREAKING PAPERS

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

A Provenance-Based Approach for Reusing Biodiversity Data Products

Daniel L. da Silva, André Batista, Cleverton Borba, Andreiwid Correa, Suelane Garcia, Pedro L. P. Corrêa
Computer Engineering Department, University of Sao Paulo, Sao Paulo - SP, Brazil
{daniellins, andrefmb, cleverton.borba, andreiwid, suelane, pedro.correa}@usp.br

Abstract - *In the last decade, several advances have been made in the publication and sharing of biodiversity data. The adoption of metadata standards and communication protocols has enabled improvements in interoperability between systems and distributed databases. However, to reuse these available data for experiments and decision-making processes, scientists need additional information about the data origins to enable its evaluation and traceability. In this paper, we propose an approach for managing provenance metadata through a computational architecture based on services and web standards. We demonstrate the application of our approach and show its practical usefulness by evaluating this architecture to manage provenance metadata generated during an ecological niche modeling. The results of our experiment show that this approach is effective in collecting and managing the provenance metadata of data products and their processes, storing these metadata in a standardized way and allowing their discovery and retrieval through the Web.*

Keywords: Data Provenance, Computational Architecture, Biodiversity Informatics, Reusability

1. Introduction

In the last decade, the international community of biodiversity informatics intensified efforts to improve processes that involve scientific data management. Several advances were achieved in the publication and sharing of biodiversity data and reuse of this knowledge for critical issues, such as conservation and sustainable use of the environment, climate change action plans, and ecosystem services [1]–[5].

The adoption of metadata standards and communication protocols allows these advances and enables improvements in data sharing and interoperability between systems and distributed databases. Among these metadata standards, we can mention Darwin Core [6], used to describe the occurrence of life on Earth and its associations with the environment; Ecological Metadata Language (EML) [7] , used to describe ecological data; and Dublin Core [8] , used to describe digital resources.

Metadata are structured information associated with an object for purposes of discovery, description, use, management, and preservation [9]. Major metadata standards have three kinds of metadata elements: descriptive elements, that describes a resource for discovery and identification; structural elements, that indicates how compound objects are put together; and administrative elements, that provides information to help management and preservation of a resource [9], [10].

These metadata standards mentioned previously have focused on the descriptive elements and some aspects of administrative and structural elements. However, to reuse these available data for new experiments and decision-making processes, scientists need additional information regarding how the files were created and updated, intellectual property rights, the original source object from which this data object derives, and technical information [10], [11]. Information that describes the data origin and processes involved in data generation and evolution are called provenance metadata.

Provenance metadata is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or an object [12]. Provenance metadata can be used for various purposes, such as evaluating the quality and reliability of data, audit processes, data versioning, reproducibility of experiments, the establishment of property data, and discovery of new data [13]. However to obtain the benefits of data provenance, scientists need to register metadata during the various stages of the data production.

Workflow Management Systems (WfMS) are the tools currently most used for documentation of scientific experiments and data processing activities. Several studies have proposed standards, techniques, and tools to improve the capture of provenance data in WfMSs and enable evaluation and reproducibility of experiments [14]–[16]. The advantages of these mechanisms are the quick configuration to capture provenance metadata and the possibility of automatic reproducibility of workflows. The main drawback is the mandatory use of a WfMS. In addition, these tools are not compatible with each other or with other tools, making it difficult to maintain provenance metadata in distributed and heterogeneous environments. Moreover, the provenance metadata generated by some WfMS are stored locally in files or internal databases, which makes difficult to discover and reuse these metadata by other scientists.

In this paper, we propose an approach to the management of provenance metadata through a

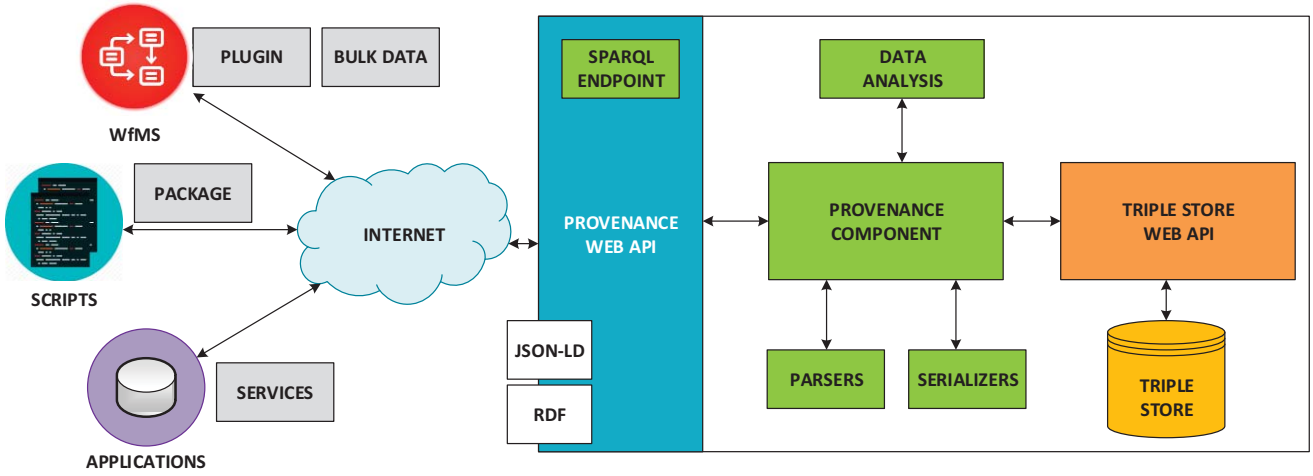


Figure 1. Overview of the computational architecture for managing provenance metadata.

computational architecture based on services and web standards.

The main contribution of this work is the novel approach used to manage the provenance metadata generated in distributed and heterogeneous environments. Our computational architecture stores the provenance metadata in a centralized database and publishes these metadata on the Web. All these operations are performed through a Web API. We also created an application profile to ensure the standardization and organization of provenance metadata in our computational architecture.

We show the application of our approach in a case study, where an implementation of this architecture managed the provenance metadata during the execution of an ecological niche modeling experiment.

2. Provenance Computational Architecture

To support capturing, recording, querying, and managing the provenance metadata we have specified a provenance computational architecture. This architecture is comprised of four main components:

- **A provenance data model** to organize and structure the provenance metadata;
- **A provenance services** that provides a Web API for recording, querying, and visualizing provenance metadata in distributed environments;
- **A provenance repository** to store provenance metadata and provide capabilities to query and analyze these data;
- **An acquisition component** to capturing and generating the provenance metadata and communicating with the provenance services.

This architecture is summarized in Figure 1, which we describe in the next sections.

2.1. Provenance Data Model

We defined a provenance data model based on the W3C PROV-DM to describe detailed information about the data and processes involved in the lifecycle of datasets and data products.

The PROV-DM was designed to describe people, entities, and activities involved in producing a piece of data or any object [12]. The concepts found in the core of PROV-DM, are shown in Table 1.

Table 1. Mapping of PROV core concepts to types and relations [17].

Concepts	Types or Relations	Names
Entity	Types	Entity
Activity		Activity
Agent		Agent
Generation	Relations	WasGeneratedBy
Usage		Used
Communication		WasInformedBy
Derivation		WasDerivedFrom
Attribution		WasAttributedTo
Association		WasAssociatedWith
Delegation		ActedOnBehalfOf

Using these types and relations, we can describe all relationships about the data objects and the processes performed in an experiment. The provenance metadata take the shape of a directed graph, considering nodes to represent the PROV Types (data, people, institutions, and

processes), and edges to represent PROV Relations (how these nodes relate to each other).

Figure 2 shows a graph that represents the relationship (derivation) between two datasets using PROV-DM. Derivation is a transformation of one entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on pre-existing entity [17].

Figure 2. In this representation of data provenance, the Dataset_B was derived from the Dataset_A.

An additional concept called Bundle is also considered. A Bundle is a named set of provenance metadata, and is itself an prov:Entity, so allowing provenance of provenance to be expressed [17]. Using Bundle, we can describe metadata about the capturing process of provenance metadata, such as the responsible scientist, the capturing tool, or additional information about this process.

Furthermore, we define an application profile based on the Dublin Core to describe additional metadata to each node (PROV-DM Types) in the provenance graph.

An application profile is a generic schema to design metadata records that meet specific application requirements, providing semantic interoperability with other applications based on globally defined vocabularies and models [18]. In the definition of our application profile, we considered the Dublin Core, Friend of Friend (FOAF), RDF Schema (RDFS) and PROV Ontology (PROV-O).

Figure 3 shows the complete provenance data model proposed in this paper.

2.2. Web API

We defined a provenance Web API, following the Representational State Transfer (REST) style [19], to provide services for recording, querying, and visualizing provenance metadata.

Resources are the fundamental concept of a REST API. A resource is an object with a type, associated data, relationships with other resources, and a set of methods that define its operations. These resources are grouped in collections and identified by an URL. The default URL schema for access the provenance services is described below.

http://**[domain]** /api/**[resources]**/**[rId]**?**bundle**=**[bId]**

Where **[domain]** is the domain name of the Web API. **[resources]** corresponds to the resources described in Table 2. **[rId]** corresponds to the Uniform Resource Identifier (URI) that uniquely identifies each resource. For Bundle association, we can use the parameter **bundle**, in the query string of URL. The use of bundle is optional.

Table 2 shows all resources provided by the provenance Web API, and their available methods.

For representation of data handled by the provenance services, we use the JSON-LD format [20]. We consider the use of JSON-LD to ensure the contextualization of terms used in the provenance data model, without the need to represent them in Resource Description Framework (RDF). Thus, we do not adding the complexity of RDF handling and new requirements to the scientific tools (client applications of the provenance Web API), which mostly are compatible with REST and JSON, but not compatible with RDF.

Table 2. Description of the resources that compose the Provenance Web API.

Resource	Description	Methods
/bundles/	The provenance of provenance. Metadata that describes the metadata capturing process.	GET POST PUT DEL
/entities/	The described data, such as datasets, records, files, scripts, and parameters.	GET POST PUT DEL
/activities/	Events that occur over a period of time and act upon or with Entities.	GET POST PUT DEL
/agents/	Responsible for an Entity or the execution of an Activity.	GET POST PUT DEL
/provenance/	Get the complete metadata about an Entity or Bundle.	GET
/bundles/<id>/data/	Get the complete metadata associated with this bundle <ID>.	GET
/sparql/	SPARQL Endpoint	GET
/entity/<id>/files/	Manage files associated with an Entity.	GET POST PUT DEL
/rdf/	Import RDF documents to the repository.	POST

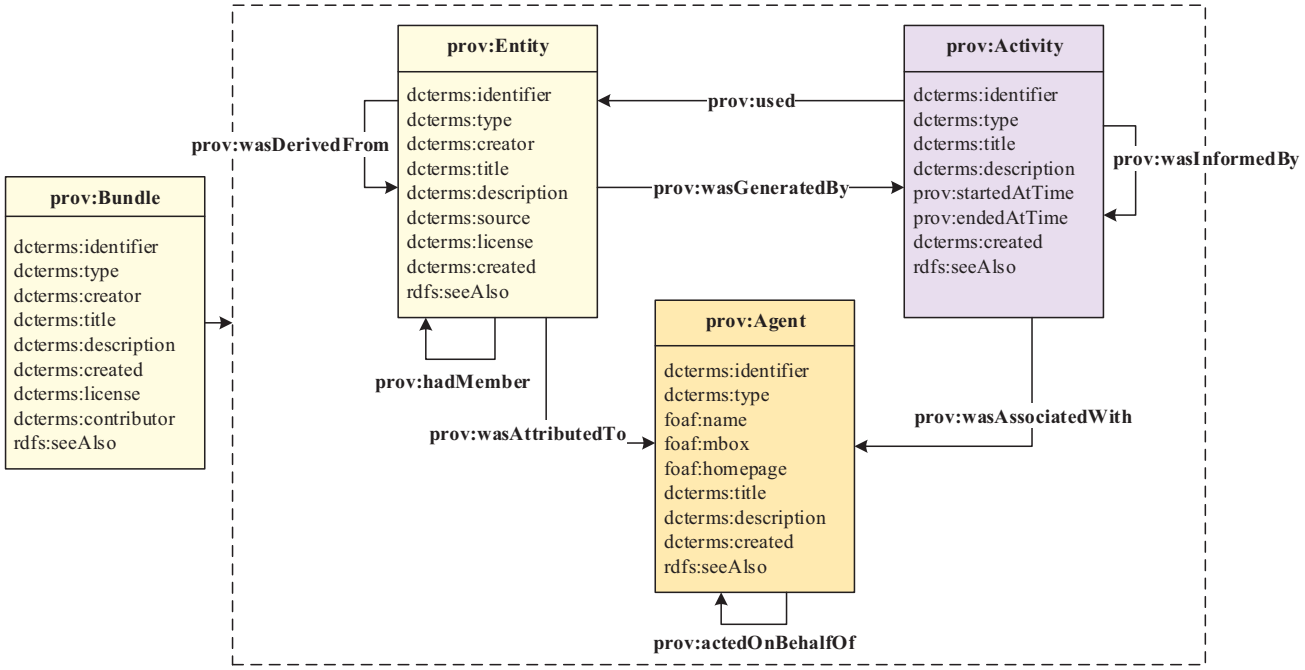


Figure 3. Proposed data model based in PROV-DM core structures and the Dublin Core application profile.

JSON-LD are compatible with JSON format and adds semantic context to a JSON document using the "@context" element. The @context element can be directly embedded into the JSON document or be an external file.

In our approach, we created an external file called prov.jsonld [1], which contextualize all definitions related to our provenance data model based on the ontologies and vocabularies considered. A part of this file can be viewed below.

```
{
  "@context": {
     "dcterms": "http://purl.org/dc/terms/",
     "prov": "http://www.w3.org/ns/prov#",
     "foaf": "http://xmlns.com/foaf/0.1/",
     "xsd": "http://www.w3.org/2001/XMLSchema#",
     "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
     "title": "dcterms:title",
     "subject": "dcterms:subject",
...
     "agent": { "@type": "@id", "@id": "prov:agent" },
     "entity": { "@type": "@id", "@id": "prov:entity" },
…
}
```

To use the provenance context with JSON documents during the web services calls, we have included a reference to this context file, which can be held in the message body or in the message header [20].

```
{
  "@context": "http://bioprov.org/static/prov.jsonld",
  "title": "Occurrences of Ziziphus joaneiro",
  "type": "Entity",
  "source": "http://gbif.org/occurences/20390190"
}
```

2.3. Provenance Repository

Although we consider the JSON-LD in provenance services, we decided to use a triple store to manage data in RDF. The main reason for our decision was to ensure interoperability with the various tools and technologies compatible with RDF, such as SPARQL. For this reason, we implemented parsers and serializers to perform the conversion of JSON-LD documents to RDF and vice versa.

The RDF is a framework for representing information on the Web. The core structure of RDF is the RDF triple. Each triple consists of a subject, a predicate, and an object. A set of such triples is called an RDF Graph [21].

The provenance metadata represented for RDF graphs are stored in a triple store. The triple store is a database specialized in storing and retrieving RDF graphs. This

[1] https://goo.gl/oPcNMD

type of database is compatible with the SPARQL[2], a RDF query language.

The provenance component, presented in our computational architecture, can access the triple store through a Web API. Many triple stores already provide Web APIs for data access. Web APIs facilitates the standardization of systems, allowing the replacement of database vendor without major impacts on the application. Moreover, it avoids problems with the triple store drivers, often outdated and limited to a few programming languages.

3. Case Study

Ecological Niche Models (ENMs) have become increasingly popular as tools for predicting the geographic ranges of species and are important for predicting changes in species distribution from past and future climatic events and for investigating patterns of speciation and niche divergence [22]. The basic premise of the ENM approach is to predict the occurrence of species on a landscape from georeferenced locality data and sets of spatially explicit environmental data layers that are assumed to correlate with the species' range [22]. This experiment aimed to analyze the ecological niche of endemic species in Brazil's semiarid region.

With the increasing availability of ecological data [23], [24], ENM has gained much attention for a wide variety of ecological applications. There are many environmental niche modeling packages and platforms available; for example, MaxEnt [25], GARP [26], openModeller [27], and so on. Existing comparisons between different niche models do not show consistent conclusions in part because the comparisons are primarily conducted on different platforms and, thus, could implement the training and testing differently. Moreover, the previous and subsequent processes, such as data preparation, data cleaning, and visualization processing, directly influence the results generated in these processes.

Therefore, there is a need to develop provenance tools that enable the detailed description of process while allowing scientists to understand details about the data, the analysis method, and the chosen experiment parameters.

This experiment aims to show the relationship between physical aspects and the occurrence of the Ziziphus joazeiro species in the Brazilian semi-arid region, through ecological niche modeling technique aided by remote sensing products.

We developed a Web API application called "BioProv" [3], based on the proposed computational architecture, to manage provenance generated during the execution of this ecological niche modeling. This application used the following technologies:

- Python[4]: a programming language;
- Flask[5]: a Python microframework;
- RDFLib[6]: a Python library for handling RDF;
- Requests[7]: a Python library for handling HTTP operations;
- AllegroGraph [8]: a TripleStore to store RDF Graphs.

Due to the popularity of R programming language in the life sciences community, we have chosen this language to create our experiment. We also implemented an R package, called "BioProv Client"[9], which provided a mechanism for the acquisition and storage of provenance metadata through the Bioprov Web API. Therefore, we can use a suite of R functions provided by the Bioprov Client to record the provenance metadata during the ENM simulation.

At the end of the experiment, we generated the reports with the provenance metadata that describes the entire experiment.

4. Results and Discussion

During the experiment execution, implemented in an R script, we used the BioProv Client functions to record the provenance metadata.

For instance, the function below records the execution of the BIOCLIM algorithm. To describe this activity, we collected the bundle of this experiment, the activity type and title, the input data and parameters, and the data generated by this activity. The complete script of the ENM experiment can be viewed at https://goo.gl/6PrcxO.

```
createActivityProvenance (bundle = bundle,
        type = 'DataModeling',
        title = 'Bioclim Algorithm',
        used = list(e_dataset, e_environmentalLayers,
                e_enmAlgorithm),
        generated = e_resultModeling)
```

The provenance metadata generated during the ENM execution was stored in the provenance repository and available on the Web. These metadata can be accessed by the Web API or using the BioProv Client functions.

The BioProv Client also enables the generation of metadata reports through the functions **getProvenanceOfBundle** and **getProvenanceOfEntity**. These reports can be generated in different formats: PROV-N, Turtle (RDF serialization), PNG, and SVG.

[2] https://www.w3.org/TR/sparql11-overview/
[3] https://goo.gl/PLk3Ie
[4] https://www.python.org/
[5] http://flask.pocoo.org/
[6] https://rdflib.readthedocs.org/en/stable/
[7] http://docs.python-requests.org/en/master/
[8] http://franz.com/agraph/allegrograph/
[9] https://goo.gl/ioiqCk

To generate the metadata report of this experiment, we have used the function below. This variable named e_**FinalresultModeling** has the identifier associated to the ecological niche map: http://bioprov.poli.usp.br/models/102.

getProvenanceOfEntity (**uri** = e_FinalresultModeling, **format** = 'png')

The ecological niche map generated by our experiment and the provenance metadata associated to this data product are available through the links below:

- Ecological Niche map: https://goo.gl/AwCEJw;
- RDF Report: https://goo.gl/tNv5cT;
- PNG Report: https://goo.gl/rVisST.

Based on the provenance metadata of this experiment, we were able to discover the raw data, details about the activities and the people involved. Thus, we could evaluate the origin of this data product and decide whether these data are suitable to be reused for my research needs. With these metadata, it is also possible to reproduce this experiment, validate its results, and gain insights about this process.

5. Conclusion and Future Work

In this paper, we proposed a provenance-based approach for reuse of biodiversity data products, supported by a computational architecture designed for distributed environments.

To validate our approach, an implementation of our computational architecture was used to manage the provenance metadata generated during the execution of an Ecological Niche Modeling. The results of our experiment show that this approach is effective in collecting and managing the provenance metadata of data products and their processes, storing these metadata in a standardized way and allowing their discovery and retrieval through the Web.

After the experiment execution, we could query and analyze the provenance metadata related to this experiment and associated with the generated data product.

We intend to carry out more case studies, capturing data provenance using other mechanisms such as WfMSs and tools like the openModeller. Moreover, we will continue the evolution of the application profile presented in this paper, including specific dictionaries of the biodiversity community, such as the Darwin Core and the EML.

Finally, we believe that due to the complexity of the provenance graphs, the graphical visualization of the provenance report is still confusing and need some improvements to facilitate the analysis of scientists.

With these efforts, we will seek to create a standardized approach for managing provenance metadata, which can be used for a variety of data analysis and data processing tools, allowing interoperability and maintenance of the provenance metadata during the entire data lifecycle.

5. Acknowledgement

This research was supported by the Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), Edital N. 005/2012.

6. References

[1] N. F. Johnson, "Biodiversity Informatics," *Annu. Rev. Entomol.*, vol. 52, no. 1, pp. 421–438, 2007.

[2] P. B. Heidorn, "Biodiversity informatics," *Bull. Am. Soc. Inf. Sci. Technol.*, vol. 37, no. 6, pp. 38–44, 2011.

[3] K. Bach, D. Schäfer, N. Enke, B. Seeger, B. Gemeinholzer, and J. Bendix, "A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research," *Ecol. Inform.*, vol. 11, pp. 16–24, Setembro 2012.

[4] J. C. Wallis, E. Rolando, and C. L. Borgman, "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLoS ONE*, vol. 8, no. 7, p. e67332, Jul. 2013.

[5] L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, and P. Pagano, "An infrastructure-oriented approach for supporting biodiversity research," *Ecol. Inform.*, 2014.

[6] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais, "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," *PLoS ONE*, vol. 7, no. 1, p. e29715, Jan. 2012.

[7] E. H. Fegraus, S. Andelman, M. B. Jones, and M. Schildhauer, "Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation," *Bull. Ecol. Soc. Am.*, vol. 86, no. 3, pp. 158–168, Jul. 2005.

[8] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin Core Metadata for Resource Discovery," RFC Editor, RFC2413, Sep. 1998.

[9] NISO Framework Working Group, "A Framework of Guidance for Building Good Digital Collections," National Information Standards Organization (NISO), Baltimore, Maryland, third edition, Dec. 2007.

[10] National Information Standards Organization (U.S.), *Understanding metadata.* Bethesda, MD: NISO Press, 2004.

[11] N. Enke, A. Thessen, K. Bach, J. Bendix, B. Seeger, and B. Gemeinholzer, "The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —," *Ecol. Inform.*, vol. 11, pp. 25–33, Setembro 2012.
[12] L. Moreau, P. Groth, J. Cheney, T. Lebo, and S. Miles, "The rationale of PROV," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 35, pp. 235–257, Dec. 2015.
[13] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Rec.*, vol. 34, no. 3, pp. 31–36, 2005.
[14] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. N. de la Hidalga, M. P. B. Vargas, S. Sufi, and C. Goble, "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W557–W561, Jul. 2013.
[15] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance Collection Support in the Kepler Scientific Workflow System," in *Provenance and Annotation of Data*, L. Moreau and I. Foster, Eds. Springer Berlin Heidelberg, 2006, pp. 118–132.
[16] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "VisTrails: Visualization Meets Data Management," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2006, pp. 745–747.
[17] K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, C. Tilmes, and L. Moreau, "PROV-DM: The PROV Data Model," *W3C Recommendation*, 30-Apr-2013.
[18] T. Baker, M. Dekkers, R. Heery, M. Patel, and G. Salokhe, "What Terms Does Your Metadata Use? Application Profiles as Machine-Understandable Narratives," *Int. Conf. Dublin Core Metadata Appl.*, vol. 0, no. 0, pp. 151–159, Oct. 2001.
[19] R. T. Fielding, "Architectural styles and the design of network-based software architectures," University of California, Irvine, 2000.
[20] Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, and Niklas Lindstrom, "JSON-LD 1.0: A JSON-based Serialization for Linked Data," 16-Jan-2014. [Online]. Available: https://www.w3.org/TR/json-ld/. [Accessed: 31-Mar-2016].
[21] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax," *W3C Recomm.*, Feb. 2014.
[22] J. D. Lozier, P. Aniello, and M. J. Hickerson, "Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling," *J. Biogeogr.*, vol. 36, no. 9, pp. 1623–1627, Sep. 2009.
[23] C. H. Graham, S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson, "New developments in museum-based informatics and applications in biodiversity analysis," *Trends Ecol. Evol.*, vol. 19, no. 9, pp. 497–503, Sep. 2004.
[24] J. Wieczorek, Q. Guo, and R. Hijmans, "The point-radius method for georeferencing locality descriptions and calculating associated uncertainty," *Int. J. Geogr. Inf. Sci.*, vol. 18, no. 8, pp. 745–767, Dec. 2004.
[25] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum entropy modeling of species geographic distributions," *Ecol. Model.*, vol. 190, no. 3–4, pp. 231–259, Jan. 2006.
[26] D. Stockwell, "The GARP modelling system: problems and solutions to automated spatial prediction," *Int. J. Geogr. Inf. Sci.*, vol. 13, no. 2, pp. 143–158, 1999.
[27] M. E. de S. Muñoz, R. D. Giovanni, M. F. de Siqueira, T. Sutton, P. Brewer, R. S. Pereira, D. A. L. Canhos, and V. P. Canhos, "openModeller: a generic approach to species' potential distribution modelling," *GeoInformatica*, vol. 15, no. 1, pp. 111–135, Jan. 2011.

Neural network for damage in a framework

Casanova-del-Angel, F[1] and Hernández-Galicia, D[2]
Polytechnic Institute National. Mexico
1: E-mail: fcasanova@ipn.mx ; www.ipn.academia.edu/FranciscoCasanovadelAngel/
2: E-mail: dhernandez@ipn.mx

Abstract
The framework used to obtain data with which the Artificial Neural Network (ANN) was developed is shown. Its geometry, properties of the material, sections of structural elements, and loads used are described. Then, the numerical model of the framework under study is developed in structural analysis software SAP2000 (Computers and Structures Inc, 2009) in order to obtain its modal parameters. In addition, a program made in Matlab® is shown, from which data with and without damage of the framework under study were obtained, and with which the ANN was developed.

Keyword: Framework, damage, neural network, static condensation and modal parameters.

Introduction

Throughout their useful life, structures accumulate gradual damage as time passes. Such damage is caused by actions and natural phenomena such as: seism's, winds, and explosions, among others. Therefore, and taking into account functionality and safety of structures, it is of paramount importance to detect damage, follow up and monitor structures in order to know their physical conditions to increase safety and structural reliability. If damage is detected on the structure early, its evolution may be observed regarding magnitude and size in order to treat it properly. Nowadays, there are many methods to detect structural damage —with advantages and disadvantages— and there is not one that may be considered the best. In this work, an Artificial Neural Network (*ANN*) is developed based on MATLAB's toolboxes (Demuth *et al.*, 1992–2009), to which condensed matrices and modal parameters were provided (frequencies, periods and vibration modes) obtained from a program made in Matlab® (Gilat, 2006), obtaining, in the end, an artificial neural network capable to detect structural damage.

Damage has been defined as loss of stiffness. In order to identify such loss, plastic articulations in structural elements of the structure under study were simulated to obtain the condensed stiffness matrix with and without damage. Then, dynamic parameters were calculated, as well as their dynamic response with and without damage, which were used to develop the *ANN*. In this work, stiffness matrices with and without damage, that is, location of damage, are known.

Method

A three-level framework was modeled in structural analysis program SAP200® (Computers and Structures Inc, 2009), from which modal parameters (frequencies, periods and mode shapes) of such framework were obtained with and without damage. A program was developed in Matlab® (Gilat, 2006), from which the global and condensed stiffness matrices were obtained at horizontal degrees of freedom, with and without damage. Modal parameters and dynamic response with and without damage were calculated, which served for the development of the *ANN*, and a failure condition was considered to define a serious damage condition.

Data with damage were obtained simulating a plastic articulation in various elements from such framework, and position of the articulation varied to obtain various damage conditions for the same framework. Data obtained were used to develop the network, and an *ANN* to detect structural damage was obtained. Damage detection was carried out in four steps: First, extraction of modal parameters and condensed matrix; second, establishing failure condition for a serious damage condition; third, treatment of modal data to be used in the development of the *ANN*; and fourth, detection of damage to be carried out with the *ANN*.

Static condensation

In order to carry out static condensation, the side stiffness matrix K_L was considered, which is associated to side coordinates of the floor, since seismic analysis of flat frameworks includes a single degree of freedom per floor. This is a stiff floor model, which only works for the analysis in view of the horizontal component of soil movement. Since the nodes are considered to be stiff, only the beams are axially stiff, and the columns are flexible, condensation is carried out as follows: the original structure has *n* degrees of freedom. The condensed stiffness matrix is obtained first, at the desired degrees of freedom, *GL*. In our case, horizontal *GL´s* are required, which are associated to displacement vector $\hat{u}$; in this case, the number of levels is identified as p and K_L is the condensed matrix of p^*p, thus reducing the total stiffness matrix to condensed matrix $K_L = k_{11} - k_{12}k_{22}^{-1}k_{21}$.

Obtaining modal parameters

To obtain modal parameters, the diagonal mass matrix is $M = [m_1\ m_2\ m_3]$. Once the condensed stiffness matrix K_L and the mass matrix M, have been obtained, the problem of own values and vectors is solved: $(K_L - \omega^2 M)\Phi = 0$, with nonsingular M and symmetric and positive defined K_L. Then, there are N real roots ω_i, being $\omega_1 \leq \cdots \leq \omega_N$, for which reason to every own value corresponds an own vector ϕ_i called own vibration modes $\Phi = (\emptyset_1, \ldots, \emptyset_n)$ where Φ is the modal matrix. To find the period, ratio $T = \frac{2\pi}{\omega}\ \forall\ \omega = \sqrt{\frac{K}{M}}$ has been used. Therefore:

$$T = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}; \quad w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & w_{N1} \\ w_{12} & w_{22} & w_{N2} \\ w_{N3} & w_{N3} & w_{NN} \end{bmatrix} (2)$$

Where T is the period and ω is frequency.

To obtain the dynamic response where there is movement on the base, the second-order linear differential equation $M\ddot{u} + C\dot{u} + K_L u = -mj\ddot{u}_g(t)$ is consdered, where u is the vector of N horizontal displacements related to the movement of the soil, $\ddot{u}$ is acceleration, and $\dot{u}$ is speed. In addition, *M, C, and* K_L are the mass matrix, the absorption matrix and the side or condensed stiffness matrix, respectively. Every value of *j* influence vector is equal to one. The right side of the differential equation may be interpreted as the seismic forces $P(t) = -mj\ddot{u}_g(t)$ which is function of time, t.

Distribution of these forces in height is defined by vector $s = mj$ and variation by time $\ddot{u_g}(t)$, which is acceleration of the land. Such distribution of forces may be expressed as summation of the distribution of modal inertial forces s_n, thus: $mj = \sum_{n=1}^{N} s_n = \sum_{n}^{N} \Gamma_n m \phi_n$, where ϕ_n is the i-th vibration mode of the structure, Γ_n is the factor of modal participation and m the mass. Seismic forces may be expressed as: $P(t) = \sum_{n=1}^{N} P(t)_n = \sum_{n=1}^{N} -s_n \ddot{u}_g(t)$

Contribution of the i-th mode to *s* is $s_n = \Gamma_n m \phi_n$ and to *P*(t) is $P(t) = -s_n \ddot{u_g}(t)$. Now, it may be seen that the response of the multiple GL system $P(t)$ is completely

in i-th mode, without any contribution from other modes. The equation representing the response of the system is: $M\ddot{u} + C\dot{u} + K_L u = -s_n \ddot{u}_g(t)$. Using orthogonal properties of modes, it may be known that none of them, except i-th mode, contributes to the response. Therefore, displacement of the floor is: $u_n(t) = \phi_n q_n(t)$, where the modal coordinate $q_n(t)$ is governed by: $\ddot{q}_n(t) + 2\zeta_n \omega_n q_n(t) + \omega^2{}_n q_n(t) = -\Gamma_n \ddot{u}_g(t)$. Consequently, $q_n(t)$ is the response of a single degree of freedom system. ζ_n is the absorption coefficient and ω_n is vibration frequency. Therefore, floor displacements for a multiple degrees of freedom system are given by: $u_n(t) = \phi_n q_n(t)$.

Program in Matlab® to simulate 2D framework

A program was developed in Matlab®, with which the framework was analyzed, obtaining data with and without damage for the development of the *ANN*. For calculation of data with damage, the total stiffness matrix K_T was obtained from stiffness matrices with articulations on one or both ends. These conditions are defined in the program in order to carry out a structural and a dynamic analysis.

The program must be supplied with the geometry and properties of material, number of floors, absorption, degrees of freedom to be condensed, *NGLC*, displaced degrees of freedom, *NGLD*, mass per floor and accelerogram. The stiffness method is used to obtain the total stiffness matrix K_T. From such is obtained the reduced total stiffness matrix *KTR*, the reduced reordered total stiffness matrix *KTRR*, and, at last, the condensed matrix K_L. Such is obtained to reduce the system in order to make a dynamic analysis to calculate horizontal displacements per floor, with and without damage. After that, the problem of characteristic values and vectors is solved to obtain frequencies and modal shapes of the framework. Finally, the response per floor is obtained applying a seism on the base of the structure. The flowchart, Figure 1, shows the sequence of steps:

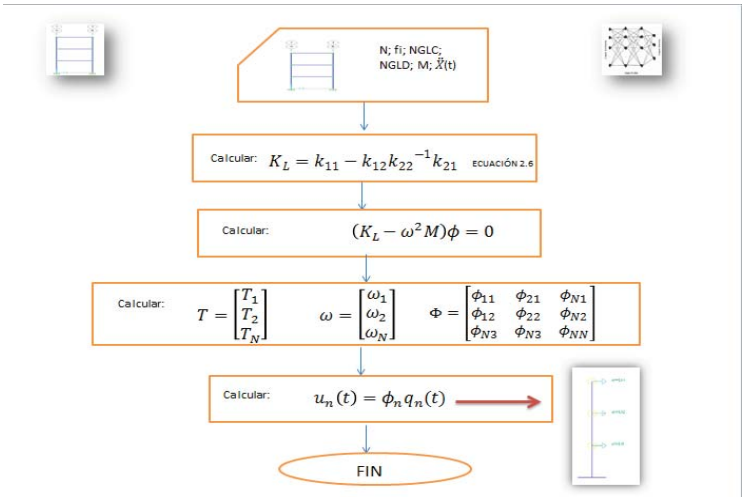


Figure 1. Flowchart to obtain data with and without damage.

Application of the neural network to detect structural damage

Let us see three cases of application of the *ANN* created. A three-level framework was analyzed, in which plastic articulations on the ends of its elements were simulated. Plastic

articulations varied in position to obtain various damage conditions. In addition, to simplify the problem, the *n* degrees of freedom system, *GL*, was reduced to a three degrees of freedom system, Figure 2, from which 48 different damage conditions were obtained to train the network. In Case 1, 16 different conditions were used, from which only one is the case without damage and the remaining 15 with damage. For Case 2, 48 different conditions were used. For Case 3, the *ANN* was trained with modal matrices from 48 cases. The failure condition described was taken into account (to determine if the structure fails or not). For every state the condensed stiffness matrix K_L, its periods, frequencies and modal matrices were obtained, which data were used to train the *ANN*.

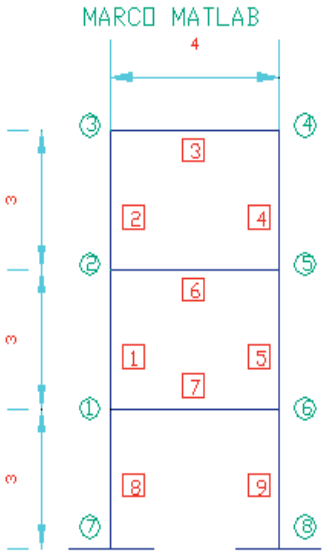


Figure 2. Flat framework to obtain training data.

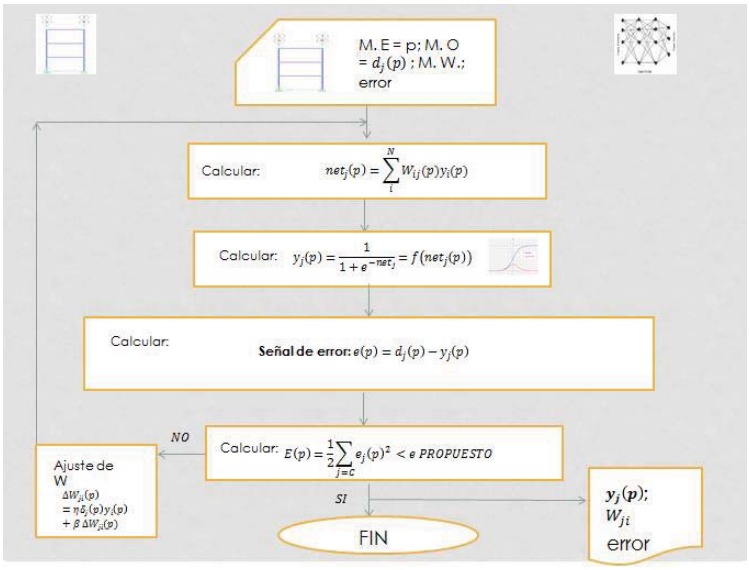


Figure 3. Flowchart of the error backpropagation algorithm or the generalized delta rule for the case under study.

Test of the *ANN*

For this example, the periods of every damage case were used to carry out the test of the *ANN*. From the above framework, a condition without damage was obtained, as well as 15 damage conditions. It must be mentioned that data were obtained with the program developed in Matlab®. Structural analysis software SAP2000® was used only to verify that data obtained in the program created in Matlab® were correct. In addition, the network used is of the backpropagation type, which uses the error backpropagation algorithm or generalized delta rule, Figure 3. The network is provided with the following data to be trained: error = 0.001 and maximum number of epochs = 10,000. From the analysis of the framework, condensed stiffness matrices K_L were obtained for every condition. Since the framework has three levels, the condensed matrix K_L is 3*3.

Results

Results obtained where as follows:

- It may be seen that maximum displacements obtained with SAP2000® are almost equal to those obtained with Matlab®, as well as the periods.

- Inputs are the weights and periods matrices and the objective. The output is a modified weights matrix and a calculated matrix close to the objective matrix. This diagram describes the error backpropagation algorithm or generalized delta rule explained above.

Damage detection

In this application, condensed stiffness matrices were used to carry out all the training of the *ANN*, different from Case 1, where the periods were used for test purposes and to verify that the *ANN* was working properly.

Conclusions

In accordance with Test 1, the *ANN* works properly calculating the objective matrix in an efficient manner, with an average 5% error. This may be lower if the number of training epochs of the *ANN* increases. From Test 2 it may be concluded that, if training cases increase, the number of epochs required for learning of the *ANN* is higher, this because input data of the *ANN* is larger, therefore, more time is required for training. Also from Test 2 it is concluded that the *ANN* has a 16% error when calculating the objective matrix, which corresponds to the increase of training cases, but if the number of training epoch's increases, the error of the *ANN* shall decrease because the *ANN* shall have more learning time. Cases of serious damage are those where articulations are in intermediate elements, and cases where damage is of the same magnitude are those where articulations are in the higher and lower corners of the framework. An articulated column on the base is more unfavorable than an articulated column in higher floors, but numerically, damage is higher when there are articulated elements in higher floors.

Acknowledgments

This article and its corresponding research was carried out, in part, with the research project IPN-SIP 20120585, and the author are very grateful to the reviewers for carefully reading the paper and for their constructive comments and suggestions which have improved the paper.

Bibliography mentioned

Gilat, A. 2006. *Matlab Una introducción con ejemplos prácticos*. Editorial Reverté. Barcelona, España. ISBN: 8429150358, 9788429150353.

Computers and Structures Inc. 2009. *Getting Started with SAP2000 Linear and Nonlinear Static and Dynamic Analysis and Design of Three-Dimensional Structures*. Berkeley, California, USA.**Howard Demuth, Mark Beale and Martin Hagan. 1992–2009**. *Neuronal Network TollBox.* ™6. User´s Guide. The Math Works Inc.

SESSION

POSITION PAPERS AND SHORT RESEARCH PAPERS

Chair(s)

TBA

ISBN: 1-60132-441-3, CSREA Press ©

ISBN: 1-60132-441-3, CSREA Press ©

The extraction mechanism of ship list and ship trajectory based on the requested region for VTS

Seung-Hee Oh[1], **Byung-Gil Lee**[1], **and Byungho Chung**[1]
[1]ICT Convergence Security Laboratory, Information Security Department,
Electronics and Telecommunications Research Institute, Republic of Korea

Abstract - *These days, we can easily listen maritime accidents in the news. It is because material, human movement is growing in the ocean has been extended to a variety of marine leisure activities. It was used for the conventional recording and playback functions of the VTS system most frequently for the analysis of maritime accident when the accident occurs. The accident occurrence time, and suspicious ship confirmed using the playback function of the existing VTS system has a big problem that it takes a lot of time. Particularly, if you do not know the exact occurrence time, such as fishing nets, damaging accidents causing damage to the ship was required a lot of time and effort to analyze the accident time. This paper proposes the mechanism that is used in post-processing maritime accident. The proposed mechanism is to quickly and efficiently identify the accident when used alone or with a conventional playback function in VTS system in a manner that the pre-extraction ship trajectory information on the accident area and the specific time condition.*

Keywords: Ship trajectory, VTS, Maritime accident, Playback

1 Introduction

The most of the import and export logistic in many countries is being carried through the sea with globalization. In particular, in the case of Republic of Korea, the 90% of import and export cargo are processed through the sea. In addition, the maritime traffic volume is increasing day by day because of an increase in the cruise ships, ferries, and leisure boats such as yachts. Despite the multilateral efforts to prevent accidents in accordance with the increase in the logistic and human traffic using the sea, various type of maritime accidents occurring at sea tend to increase [1][2].

Among the statistics of damage to the ship, the ship accident that occurred caught in fishing nets during navigation is known as a trend of increasing year by year [3]. As with all types of accident, if the maritime accident occurs, information gathering in order to understand the responsibility and the cause of the accident takes place preferentially.

Maritime accident investigation is achieved by utilizing the ship is installed on the equipment that a black box voyage data recorder (VDR) for collecting navigation data or information utilized the trajectory of a ship that is stored and managed in the Vessel Traffic control Service (VTS) system [4].

It is clear the accident occurred hours in the case of conflict accident between the ships. In this case, the range of gathering accident information is limited. However, the case of fishing nets damage incidents are difficult to determine precisely the time the accident occurred. Therefore it becomes vast amount of information collected to find the actual attacker/offender.

In this paper, we provide the mechanism which solves the drawback of requiring a lot of time to identify an accident in existing VTS system. The proposed mechanism by working rather than simply save the track information of the ship to be collected in the VTS, and help to quickly and effectively grasp the contents of the time of the accident and the accident situation.

This paper is organized as follows. In section 2, we introduces existing method of analysis maritime accidents. Section 3 proposes trajectory extraction mechanism for analysis of maritime accidents and finally, section 4 gives conclusion.

2 The existing method of analysis maritime accidents

Figure 1 shows the way to handle the accident analysis about maritime accidents. If a maritime accident occurs, it will report the VTS Center or coast guard (Step 1). To investigate the accident, data collection on the maritime accident period is carried out (Step 2). After collecting related data, analysis of collected data is made (Step 3) and then find out the accident time and suspicious ship of the maritime accident (Step 4). The legal and ethical process for the maritime accident is underway (Step 5).

In general, the step 2 and step 3 in Figure 1 are required a lot of time to find out the accident. In this paper, we propose a way to shorten the time required for data collection and analysis of data relating to incidents of step 2 step 3 [5].

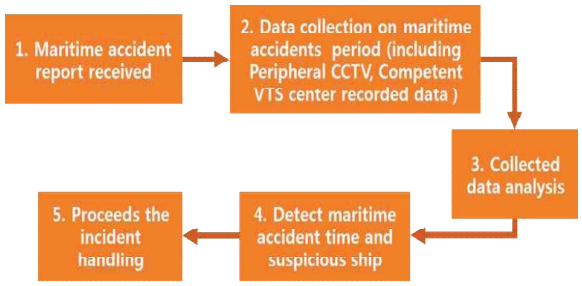


Figure 1. Existing maritime accident analysis procedure

The playback functions of the VTS system is a method to determine the time of the accident and incident information, importing historical data stored within the maritime traffic control systems at a faster rate than the actual play. Figure 2 shows the popup window of playback function GUI in existing VTS system (NorControl VOC).

However, the existing method has a disadvantage in that it is present requires a lot of time to determine the exact time of an accident, even if the speed reproduction by applying in most cases, except when exactly identified the maritime accident occurrence time.

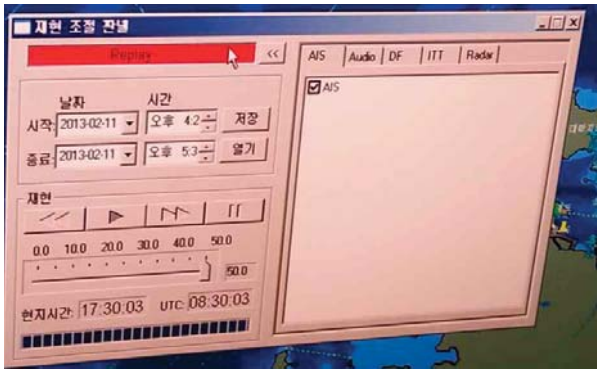


Figure 2. NorControl VOC Playback popup window

If you recognize an accident damaged fishing nets approximately a week later, this time as shown in Table 1 are required in order to confirm the suspect ship and the exact time of the incident in the existing VTS system.

Table 1. The processing time to confirm the playback results (A week source data reference)

Playback Speed	1 Speed	10 Speed	20 Speed	30 Speed	50 Speed
Analysis Time	7 days	16hours 48min.	8hours 24min.	5hours 36min.	3hours 22min.

Even applying the maximum playback speed of 50 access time speed playback of the current VTS system takes the time of at least 3 hours 22 minutes for the accident analysis.

Moreover, in reality, it is very difficult to check visually verify when executing the recording playback speed increased by more than 10 speed. Therefore it takes more time to the real event analysis in current VTS system.

3 Proposed trajectory extraction mechanism for analysis of maritime accidents

In this paper, we propose the mechanism to quickly find out the maritime accident time. It is the mechanism of extracting the trajectory information in a way that can be utilized in the previous step to determine the accidents at sea through a conventional playback.

In the case of fishing nets damage accident, it is possible to grasp the location information that has been installed of fishing net. The proposed trajectory information extraction mechanism has an advantage of offering to select a desired region other than the control information for the entire area. That is, it is possible to shorten the extraction information and verification time, because only selected desired area.

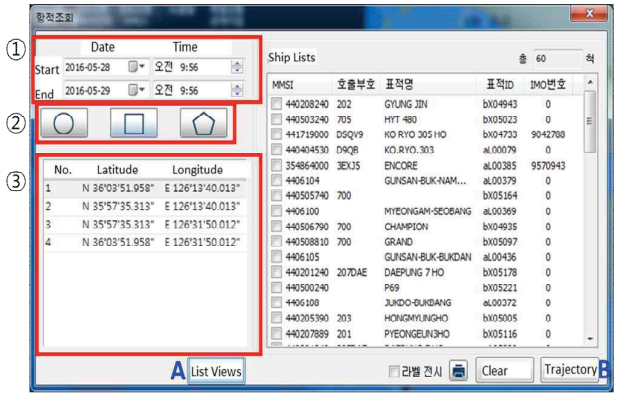


Figure 3. Trajectory query popup window

Figure 3 shows a pop-up window for area-based trajectory query. First, select the shape to be displayed in the area to request in a pop-up window with the start date and end date (①). Then, select the shape for drawing, directly drag on the electronic chart in the VTS operating system drawn by selecting the desired area (②). The shapes for area-based trajectory query can be selected from circles, rectangles and polygons. Draw the shape of the electronic chart, latitude and longitude coordinates are displayed as ③ in Figure 3. After you select start date, end date and desired area, press the

button “List Views” (A) to request trajectory list. It is provided with a list of ships which request time there is a past history of the desired area.

Select one or more ships appeared on the ships list window, and then click “Trajectory” button (B) is that the ship is passed by trajectory appear as red dots as shown in Figure 4.

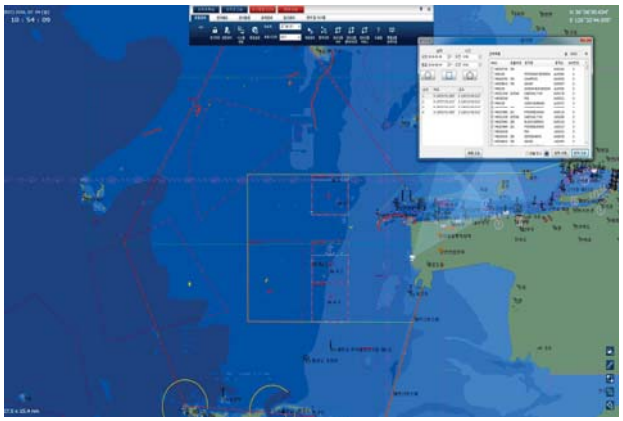


Figure 4. Display ship trajectories in VTS operating system

This is displayed on the electronic chart of the VTS operating system and even if it is not used for playback functionality through which it is possible to extract the suspicious ships. You can also check suspicious ship accidents time through previous trajectory the area.

The proposed mechanism can be confirmed by checking the display directly to a suspicious ships by ships trajectory. When using the proposed mechanism to extract information using ship trajectory in accident area with playback function the manner of maritime accidents incurred after the first time identified the maritime accident it is more effective. This approach is in progress the analysis of maritime accidents rapidly, it can be saved through human resource as well as time resources.

The time it takes to receive the results of the analysis ships list passage during a week for the particular area is represented differently depending on the regional maritime traffic control through traffic volume characteristics. In this paper, we utilize the recorded data collected as data that simulates a virtual target Gunsan VTS center in Republic of Korea.

Table 2 shows the example of request condition for ship list in the selected area in Gunsan VTS center. Applying the start date, the end date, and the area shape information for the test in this paper, is as follows. The ZoneType 3 means shape of Rectangle. Table 3 shows the time required to extract a list of ships to analyze the passage for a week in the selected area.

The request processing time is always different depending on the number of stored trajectory, the number of stored ships, and the size of selected area.

Table 2. Request condition for ship list in the selected area

```
== [2016-07-04 01:17:37] ==  [UTC]
Request Period:
[2016-05-28 00:56:43.0] ~ [2016-05-29 00:56:43.0]  [UTC]
ZoneType [3] -> means Rectangle
Index [1], Latitude [36.064433], Longitude [126.227781]
Index [2], Latitude [35.959809], Longitude [126.227781]
Index [3], Latitude [35.959809], Longitude [126.530559]
Index [4], Latitude [36.064433], Longitude [126.530559]
```

Table 3. Request processing time of ship list in the selected area

Request period	Analysis and extraction time of ship list(seconds)	Result ship #	Processing trajectory #
1Day (2016/05/28~2016/05/29)	3.588	60	18,742
2Days (2016/05/28~2016/05/30)	40.202	1,148	494,569
3Days (2016/05/28~2016/05/31)	47.845	2,354	997,063
7Days (2016/05/28~2016/06/04)	464.787 (7min. 44sec.)	10,815	14,142,719

To apply the proposed mechanism to determine if a maritime accidents can be confirmed that a much shorter time than when applied to only the existing playback method. The scheme proposed is requested by the ships trajectory information using an extracted lists of the ship. This can approximately confirm the suspicious ships 120 times faster or more compared to the 10-speed access time playback method.

Our mechanism is possible to help to reduce human resources and time resource to find the ship that caused the accident using information collection about the ship was transiting the area, and also helpful to narrow the range of the ship which is suspected to additional request the ship trajectory based on the collected information.

4 Conclusions

In this paper, we are dealing with area-based trajectory extraction mechanism that can be applied to identify the suspect ship and maritime accidents incurred in the event of an accident, such as fishing nets damaged.

The proposed mechanism can complementary to disadvantage that it takes a lot of the way things used an existing playback function of time, quickly and efficiently determine the time the accident and suspect ships and is very helpful in resolving maritime accidents.

The proposed mechanism includes a preprocessing for storing the location information of ships by parsing the received periodically. If the request is occurs, it uses the preprocessing in a way that extracts the location information data of area-based trajectory. The proposed mechanism is working further progresses faster for the study results for the preprocessing mechanism and providing an index value applied to the request.

5 Acknowledgement

This work was supported by ETRI through Maritime Safety & Maritime Traffic Facilities Management R&D Program of the MPSS (Ministry of Public Safety and Security)/KIMST (2009403, Development of next generation VTS for maritime safety).

6 References

[1] Roman Smierzchalski, Zbigniew Michalewicz, "Modeling of ship trajectory in collision situations by an evolutionary algorithm", IEEE Transactions on Evolutionary Computation, Vol. 4, pp227-241, Sep. 2000.
[2] Feixiang Zhu, "Mining ship spatial trajectory patterns from AIS database for maritime surveillance", 2011 2nd IEEE International Conference on Emergency Management and Management Sciences (ICEMMS), pp772-775, Aug. 2011.
[3] "Accident damage fishing nets attention needed", http://www.haewoon.or.kr/ksa/bbs/selectBoardArticle.do?nttId=17659&bbsId=B_000735&menuNo=700039&viewType=, Korea shipping association.
[4] J.Y.Jeong, "Study on the role VTS", The Korean society of marine environment and safety, 2014 Spring conference, pp236-238, June 2014.
[5] Seung-Hee Oh, JoongYong Choi, Kwantae Cho, Byung-Gil Lee, "The Efficient Trajectory Extraction Mechanism for Maritime Accidents", Korea Information Processing Society 2015 conference, Vol. 22, No. 2, pp30-32, Oct. 2015.

A Dynamic Model for Quality Control in Crowdsourcing Systems

Reham Alabduljabbar[1] **and Hmood Al-Dossari**[2]
[1] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, ralabduljabbar@ksu.edu.sa
[2] Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, hzaldossari@ksu.edu.sa

Abstract - *Crowdsourcing is an increasingly popular approach for utilizing the power of the crowd in performing tasks that cannot be solved sufficiently by machines. However, the challenge of effectively exploiting such systems lies in the quality of the obtained outcome from the crowdsourcing workers. While many studies and techniques have been conducted to manage the quality, they have mostly focused on managing the quality statically and used the same quality control mechanism for evaluating different types of tasks. Nonetheless, to obtain high-quality results, different quality control mechanisms should be applied to evaluate the different type of tasks. Hence, in this paper, we present our ongoing research in the domain of quality control in crowdsourcing systems and propose a task ontology-based model to facilitate identifying the most appropriate quality control mechanism for a given task.*

Keywords: *Crowdsourcing; Quality Control; Task Ontology; Reputation; Framework*

1 Introduction

Technology has emerged to accomplish many tasks automatically in order to save human's effort, time, and cost. However, not all tasks can be performed by machines, and human involvement might be required to help in conducting such tasks. In other words, humans may outperform machines in accomplishing some tasks that require very basic human skills, especially those tasks related to creativity, natural language processing and image understanding [1]. The need to solve such kind of tasks has raised a new computation model, called Human Computation. Luis von Ahn in [2] defines human computation as: *"a paradigm for utilizing human processing power to solve problems that computers cannot yet solve*". Broadly, human computation covers a wide diversity of applications such as Games with a Purpose (GWAP), Human Sensing, and Crowdsourcing.

Crowdsourcing introduces a new way for organizations and individuals to leverage the humans' knowledge and intelligence towards accomplishing special tasks that are difficult to fulfill effectively with machines alone. For example, the crowd may be invited to tag a photo, translate written text, transcribe an audio file or perform usability testing. This can help organizations to extend their resources, improve productivity while reducing costs and minimizing time [3].

Crowdsourcing applications have introduced a motivating pool of data for researchers. For example, crowdsourcing is often used to collect data in smart city projects as it provides a cheap and easy way to access data [4]. Moreover, crowdsourcing has been considered for idea generation and idea selection in the context of smart city innovation [5]. However, due to the openness nature of crowdsourcing that allows anonymous people to participate, and introduces erroneous and malicious contributions [6], many key challenges need to be considered. These challenges include, but not limited to: data analysis and processing, task design, task routing, privacy and ensuring data quality. This paper is concerned with the quality control issue and evaluating the trustworthiness of the collected data.

Different approaches have been proposed in the literature to control the quality in crowdsourcing systems such as in [7]. However, these approaches maintained the quality statically and targeted a particular domain. In other words, as illustrated in [8], a mechanism that works well for some tasks might work poorly for another one. For example, a QCM that is suitable for evaluating the outcome of an essay writing task is not appropriate for evaluating the outcome of an image labeling task.

In this work, we present a task ontology-based model that can be utilized to identify which QCM is most appropriate for a given task. To do so, we suggest enriching the proposed ontology by historical information that has been collected from previous tasks. Such information may expose a valuable knowledge that can assist in determining the most appropriate QCM.

2 Quality control issues in crowdsourcing

A major limitation in current systems is their reliance on one quality control approach. That is, the same QCM is used for different types of tasks. As highlighted in [8], requesters cannot customize the quality control approach based on their needs. Besides, requesters lack the knowledge about what is the best quality control approach that will generate the best results of their tasks. Another limitation that has a direct influence on the quality of the result is workers' reputation as it has been used as a metric for evaluating the quality of their responses.

However, current approaches calculate worker's reputation in general without taking into account the type of the accomplished task. That is, a worker might gain a high reputation from correctly accomplishing image labeling tasks, but this does not imply he or she can perform well in a text-translating task. In this work, we attempt to address these limitations.

3 Proposed model

Our proposed model (Figure 1) has the following key features that distinguish it from other work in this area: 1) The ontological representation of tasks and QCMs; 2) The dynamic identification and mapping of the suitable QCM to a given task; 3) The provision of a reputation system for the QCMs themselves rather than a reputation system for the workers.

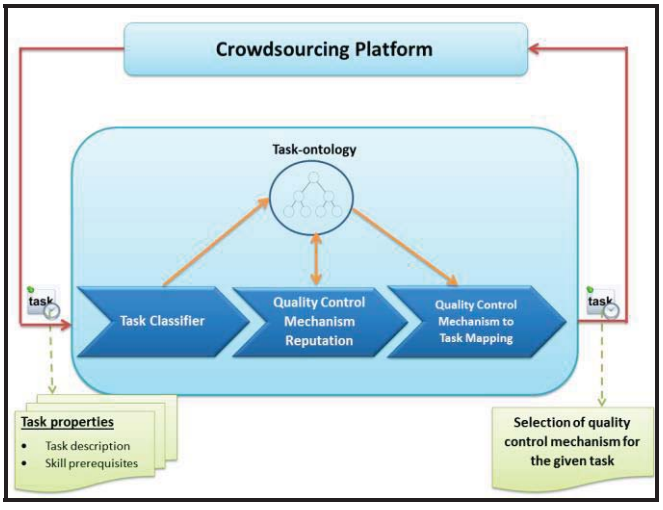


Figure 1 . High-level presentation of the proposed model.

The proposed model consists of the following components:

1) *Task Classifier:* handles task identification by classifying and clustering tasks. For example, in Amazon Mechanical Turk, one of the most popular crowdsourcing platforms, tasks are described by tags. The Task Classifier starts by fetching a task from AMT then the tag extraction process extracts information such as: title, description and required skills. Next, each task will be aligned to a corresponding existing or similar ontological instance in the system by measuring similarities with other instances. The ontology instances are hierarchy-organized based on Task Type. Classifying tasks based on their types and annotating them with tags will provide additional semantical information by mapping ontology instances as descriptors to a certain Task Type. This helps in discovering relations between tasks and facilitates the identification and recommendation of the best QCM. Algorithm 1 clarifies the classification process.

Algorithm 1 : Task Classification

1: **input**: Task *T* from a crowdsourcing platform

2: Extract all tags and keywords in *T*
3: Apply Pre-processing step (e.g. stemming and stop-word cleaning)
4: Lookup the tags instances in the Task Ontology
5: Measure similarity
6: Add *T* to the most similar Task Type as an instance
7: Add the extracted tags as instances to describe *T* and the Task Type

2) *Task Ontology:* The ontology structure is chosen for tasks representation for the following reasons:

- Tasks are the core element in any crowdsourcing systems, and they can be classified into different categories (e.g. NLP tasks, Image processing tasks, etc.).
- The deployment of the task ontology will facilitate the structured capturing of data related to crowdsourced tasks and details about their evaluation mechanism in an organized and meaningful way.
- This representation can be adopted and reused within crowdsourcing systems.

3) *QCM Reputation System:* Two aspects have to be considered when assigning a QCM to a given task. The first one aims at identifying the level of QCM suitability to the nature of the given task, and the second one derives the requester satisfaction on the returned result. The QCM Reputation Engine component handles the calculation of QCM's reputation score, and the level of suitability is computed based on historical data describing the performance of the QCM with this given task. Based on the above aspects, a reputation score is calculated for each QCM associated with the requested task. Next, a ranked list of QCMs is created, and then, the list is passed to the mapper to select one QCM from this list. Algorithm 2 describes the functioning of the ranking algorithm.

Algorithm 2 : QCMs ranking algorithm

1: **input**: Task T to generate a ranked list of QCMs used previously to evaluate T.

2: Create a list L of all $Q \in$ QCMs that has been used previously to evaluate T.
3: Calculate reputation score for every Q in L:
4: **for each** Q_i in L **do**
5: $scoreQ_i \leftarrow 0$
6: **for each** Q_iRate_j in Q_i **do**
7: $scoreQ_i = scoreQ_i + Q_iRate_j$
8: **end for**
9: $scoreQ_i \leftarrow scoreQ_i/j$
10: **end for**
11: Rank Q in L by reputation score.

12: **output:** Ranked list of QCMs used with T.

4) *QCM-Task Mapper:* With the utilization of the ontology semantics and the output of the reputation algorithm, a mapping algorithm is proposed to enable the automation of mapping the suitable QCM to the crowdsourced task.

4 Conclusions

In this paper, we have presented a high-level presentation of a task ontology-based model, which can be utilized by the crowdsourcing systems to identify which QCM is most appropriate for a given task. Our future work centers on the implementation of the model and its components. Furthermore, a simulation framework will be implemented for evaluating the model and its components.

5 References

[1] Lesandro Ponciano, Francisco Brasileiro, Nazareno Andrade, and Lvia Sampaio. Considering human aspects on strategies for designing and managing distributed human computation. Journal of Internet Services and Applications, 5(1), **(2014).**

[2] Luis Von Ahn. Human Computation. PhD thesis, Pittsburgh, PA, USA, **(2005).**

[3] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on mechanical turk. Behavior Research Methods, pages 1–10, **(2014).**

[4] L Cilliers, S Flowerday, Information security in a public safety, participatory crowdsourcing smart city project, World Congress on Internet Security (WorldCIS-2014), **(2014)**

[5] Schuurman, D., Baccarne, B., De Marez, L. and Mechant, P., Smart ideas for smart cities: investigating crowdsourcing for generating and selecting ideas for ICT innovation in a city context, Journal of theoretical and applied electronic commerce research, 7(3), Pages 49-62, **(2012)**

[6] Kanhere, Salil S. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. Mobile Data Management (MDM), 2011, 12th IEEE International Conference on. Vol. 2. IEEE, **(2011).**

[7] Nuno Luz, Nuno Silva, and Paulo Novais. A survey of task-oriented crowdsourcing. Artificial Intelligence Review, pages 1–27, **(2014).**

[8] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H.R. Motahari-Nezhad, E. Bertino, and S. Dustdar. Quality control in crowdsourcing systems: Issues and directions. Internet Computing, IEEE, 17(2):76–81, March **(2013).**

Design and Implementation of Twitter Based Food Recommendation System

SungHo Kim [1], **YongSung Kim** [2]
[1, 2]Division of Computer Science & Engineering, Chonbuk National University,
567 Baekje-daero, deokjin-gu, Jeonju City, 54896 South Korea

Abstract - *In this paper, we proposed a system offering the Twitter user help in individual decision. The knowledge base in this paper is designed and implemented for Twitter real-time retrieval system. Twitter is a free social networking. Twitter posts microblog that is different from the long blog written seriously, so it is very convenient to use on mobile phones and other mobile terminals. Twitter users exchange information timely and efficient with the short text, and it is possible to exchange with user in other network platforms by using of the related links. The collected 'tweets' are classified by food types and a table of food types and corresponding key words such as sensorial adjectives, weather adjectives, anniversary and so on. At last, the table and weight values are stored in the ontology so that it gives users recommendations like a professional.*

Keywords: collective intelligence, retrieval system, sensorial adjective

1 Introduction

Today, as internet and online space arise, people experience a new space of communicating which is very different from the traditional one. By internet, people share their experiences and interests online without the restrictions of time or space. For example, we can chat by e-mail with the friend on the other side of earth like he right beside us.

Many people choose Web or Social Network Service (SNS) to share their experiences. People with a same concerns obtain comments and opinions from others. After the collection and analysis of these information, it will be a practical information for people's decision [1].

In this paper, our objective is by expanding the SNS including Twitter users' personal social activities such as individual opinion and evaluation to propose a system that helps the SNS users make individual decisions. We design and implement a food recommendation system to achieve this objective.

To implement the proposed system, first, a large amount of 'tweets' with diet related terms are collected by Twitter API. Then, the collected 'tweets' are classified by food types and a table of food types and corresponding key words such as sensorial adjectives, weather adjectives, anniversary and so on. At last, the table and weight values are stored in the ontology so that it gives users recommendations like a professional.

2 Related Work

2.1 SNS and Twitter

Recently, the representative SNS in Korea are mini-homepage Cyworld, Twitter, Facebook, Myspace and so on. Among them, Twitter is a free social networking as microblog service. The 'tweet' can be transmitted within 140 words by SMS, Instant Messenger, E-mail etc. These 'tweets' are posted on users' profile page and browsed by others. Twitter support multi-language and no geographical residence by which users communicate with short texts like face to face.

Because on Twitter, the microblog can be write freely, it is easy to upload the text in by mobile communication terminal. At the same time, with the widely use of mobile devices, the usage of Twitter is increasing rapidly. In Korea, the usage of Twitter is also growth explosively because of the perfect construction of internet service infrastructure and the increase of mobile devices.

The Twitter users' personal social activities such as individual opinion and evaluation can be expanded as collective experiences which have influence on individual decisions in turn.

2.2 Semantic Information and Ontology

Intelligent Soft Agents proposed by Tim Berners Lee is the web environment that computer generates new information with the application to existing information on web. In this web environment, searching information and making a decision are all done by computer automatically. It means that a paradigm which computer is able to extract meaningful information automatically by the combination of concept of web-text metadata and semantic information. The combined information should be expanded and shared [3].

On the semantic web, there is a relationship between knowledge representation and information resources. Or we can call this relationship 'connectivity'. The implementation of web is based on this 'connectivity', and web becomes a distributed information space because of 'connectivity'. The connectivity on traditional web is that by hyper-text link the locations on virtual space are connected. But, on the semantic web, it stresses the semantic connection among information resources. That is to say, with the metadata of human, the information can be processed with 'meaning in brain' coded by machine program. The unit of meaning that machine is able to analyze should be defined [2].

The concept from people's discussion about what they see, hear, feel and think can be model into a form which computer understands. This model with explicit application conditions is called Ontology. The aim of semantic web is to develop the standard and technology that make the information on web be understood by computer and are used to support semantic search, data integration, navigation and office automation. Ontology contains inference rules and the relation for knowledge representation in the web-text that make computer processes information intelligently. By this way, the information users request can be supported exactly. So the conceptualization and specification of the knowledge in a particular field are done by Ontology for semantic web. The definition of Ontology is the conceptualization and the relation representation of the common knowledge of human and computer.

So Ontology is a kind of database which is different from the usual relational database. Ontology has the hierarchical structure, other relationships and restrictions of concepts. When we search 'travel by ship', first, 'travel', 'by', 'ship' are extracted. And then the relation of phases 'ship<sea<transportation<by' is performed. Only the phases have relation with them are shown and others are hided, as shown in Fig 1.

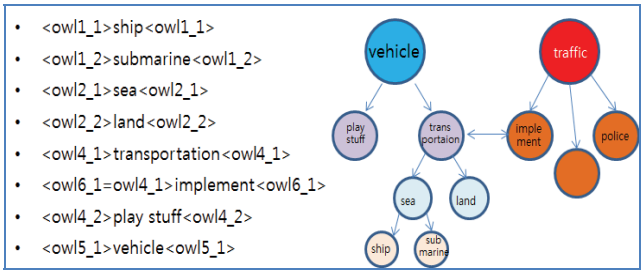


Fig 1. Implementation of Ontology and search

All illustrations, drawings, and photographic images will be printed in black and white. We recommend that you examine a printed copy of your paper (in black and white) and make the final adjustments before submission. All illustrations must be numbered consecutively (i.e., not section-wise). Center the figure captions beneath the figure. Do not assemble figures at the back of your article, but place them as close as possible to where they are mentioned in the main text. Figures can span the two columns if need be within the page margins.

3 Food Recommendation System

In the food recommendation system based on Twitter, the information collection and factor analysis are implemented by API system that is able to access Twitter. A database is built with the collection on-line information is by off-line collectivizing system. The database is connected with the recommendation system so that it can answer the questions from users. With the interaction among Twitter, database system and users' questions, the database of this recommendation system is improved like Fig 2.

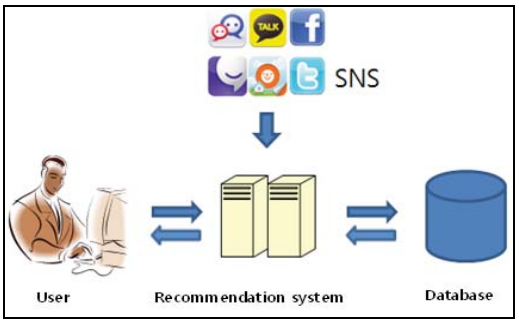


Fig 2. Flow chart of food recommendation system

3.1 Process of recommendation system

The process of the food recommendation system is shown as Fig 3. This recommendation system has the feature of feedback system.

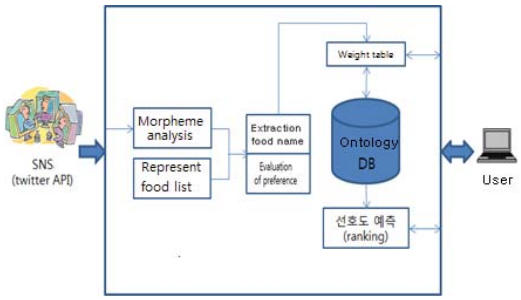


Fig 3. Process of recommendation system

3.2 The list of represent food

The name of food recommend is obtained by the list of represent food and the morpheme analysis using the extracted information from 'tweet' by API. The evaluation of preference is applied in the morpheme analysis part. The evaluation is based on the number of related words for the extracted food. After the decision of recommended food type and preference, a table of weight value is generated as an Ontology database[6].

The file of 'tweet' that collected by Twitter API is stored. The food name, weather, feeling, anniversary and other word are extracted from the file by the code below.

```
* Extraction of food name
File.open("Menu.txt").read.split("\n").each do |line|
  menu.push(line)
  myhash[line] = [0] * 51 # The number of food related word
end
* Extraction of weather, feeling and anniversary in stored tweets.
File.open("option_list.txt").read.split("\n").each do |line|
  menu1.push(line)
end
```

3.3 Weight value table

The 'tweet' contains food name, weather, feeling, anniversary is stored as a line. After that, the weight value table is generated by frequency analysis. The process of weight value table generation is shown as below.

```
* Weight value table
#write_dir = "./"
tweet_file = File.open("1_result_weight.txt", "w")
* Open the stored 'tweet' and store the frency
File.open("0_result.txt").read.split("\n").each do |line|
  # Extract the words satisfied conditions and store the frency
  for ind in 0..(menu.size-1)# process for all foods
    if line.include?(menu[ind])
      for j in 0..(menu1.size-1)
        if line.include?(menu1[j])
        # Input and store the option words on the menu
        myhash[menu[ind]][j] =myhash[menu[ind]][j].to_i + 1
        end
      end
    end
  end
end
# file of weight value table
tweet_file.puts myhash
tweet_file.close
```

3.4 Building the ontology

The ontology using sematic web is built to compare the vocabularies extracted by morpheme analysis on Twitter and the one from users' questions. The sematic web is implemented by building the ontology with a large amount of vocabularies. The ontology usually gathers class, property and constraints. The domain of ontology and concept category are decided and the taxonomy among classes is set up [5].

The stored represent food name and related word of food obtained from Twitter are collected by morpheme analysis and are stored as the data of ontology. The concept of thesaurus is important for ontology building. It is because the well-built thesaurus can be transformed to ontology easily. According to this view, in the OWL proposal of W3C, the using of OWL Lite support the function suitable for web application and utility for ontology transformation[4].

In this paper, the ontology is built by XML language. Instead of building a completed ontology, with simple programming an ontology structure is built by text analysis and weight value table. Beside, an actual inference system is not generated.

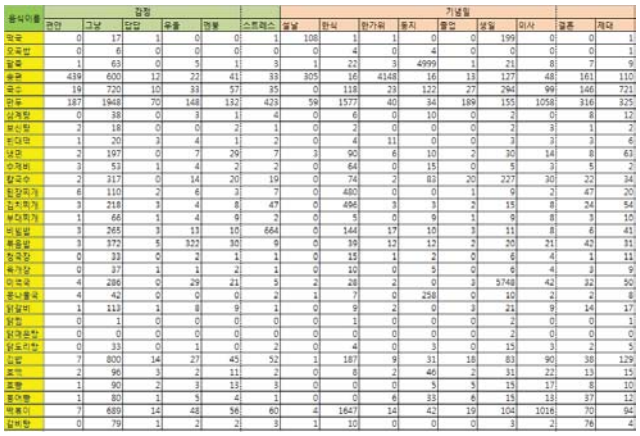
At the present stage, the ontology is not a completed ontology structure containing inference system. We just propose the possibility that a thesaurus based ontology can be built by the application of collective intelligence model with Twitter and SNS. The ontology designed in this paper shows the collective model can be used to build an ontology for different kinds of recommendation systems.

4 Experimental Results

Fig 4. and Fig 5. are the weight value table outputted by excel. The weight values pulse 1 when the word of weather, feeling, anniversary, etc. occur in the 'tweet' containing the food name.

음식이름	날씨								감정			
	비	눈	천눈	바람	폭염	봄	여름	가을	겨울	즐거움	설렘	행복	상쾌
떡국	87	8	0	4	0	31	21	15	24	0	0	1	0
오곡밥	1	1	0	0	0	0	0	0	0	0	0	1	0
갈비	278	83	0	42	0	16	28	21	36	0	0	53	0
잡채	3022	660	0	205	0	80	40	228	45	25	2	11585	6
국수	9153	499	9	196	211	147	421	334	129	15	2	280	11
냉수	8929	1705	3	361	0	180	48	326	257	77	14	410	2
삼계탕	507	34	0	11	0	27	72	39	21	0	0	16	0
보신탕	91	38	0	7	0	20	26	6	6	0	0	4	0
빈대떡	387	26	0	8	0	7	2	6	0	4	0	20	1
냉면	3055	129	1	21	2	86	220	109	132	3	1	25	0
구재비	4437	38	0	38	0	8	9	23	6	0	0	19	0
칼국수	3260	233	2	118	0	42	46	116	68	10	2	121	2
된장찌개	1071	57	0	12	0	13	7	16	11	4	1	36	1
김치찌개	1477	103	0	14	0	22	8	29	16	3	1	87	4
부대찌개	761	47	0	12	0	9	1	7	5	2	0	26	0
비빔밥	16501	118	0	61	0	53	31	623	30	1	0	39	0
볶음밥	2168	150	0	15	0	383	10	27	6	1	0	69	3
쌈국장	405	22	0	17	0	13	6	16	7	1	0	9	2
육개장	461	31	0	7	0	6	5	8	9	0	0	10	0
미역국	1079	174	0	28	0	27	8	15	13	1	1	570	2
콩나물국	473	52	0	2	0	5	0	23	3	0	0	6	1
갈갈비	17953	87	1	18	0	37	7	25	14	1	0	112	0
유정	64	2	0	0	0	0	0	0	0	0	0	0	0
닭개장탕	4	0	0	0	0	0	0	0	0	0	0	0	0
감도리탕	301	17	0	9	0	0	0	4	0	1	0	13	0
김밥	4226	692	1	126	0	111	31	214	26	7	7	242	12
포백	842	90	1	121	1	27	36	60	175	0	0	26	1
포뽕	348	198	3	53	0	32	92	167	356	0	1	9	0
뽀어뽕	505	196	2	40	0	30	71	139	360	2	0	59	2
떡볶이	5178	1248	0	129	0	151	31	154	141	5	3	356	2
갈비탕	2776	42	0	11	0	23	4	9	1	0	0	18	0

Fig 4. Weight value table 1 outputted by excel



음식이름	감정						기념일								
	관심	그냥	답답	우울	멘붕	스트레스	설날	한식	한가위	동지	졸업	생일	미사	결혼	제대
막국	0	17	1	0	0	1	108	1	1	0	0	199	0	0	1
오곡밥	0	6	0	0	0	0	0	4	0	4	0	0	0	0	1
팥죽	1	63	0	5	1	3	1	22	3	4999	1	21	8	7	9
송편	439	600	12	22	41	33	305	16	4148	16	13	127	48	161	110
국수	19	720	10	33	57	35	0	118	23	122	27	294	99	146	721
라면	187	1948	70	148	132	423	59	1577	40	34	189	155	1058	316	325
삼계탕	0	38	0	3	1	4	0	6	0	10	0	2	0	8	12
보신탕	2	18	0	0	2	1	0	2	0	0	0	2	3	1	2
비빔면	1	20	3	4	1	2	0	4	11	0	0	3	3	3	6
냉면	2	197	0	7	29	7	3	90	6	10	2	30	14	8	63
수제비	3	53	1	4	2	2	0	64	0	15	0	5	3	5	2
칼국수	2	317	0	14	20	19	0	74	2	83	20	227	30	22	34
회장떡가	6	110	2	6	3	7	0	480	0	0	1	9	2	47	20
백치떡가	3	218	3	4	8	47	0	496	3	3	2	15	8	24	54
부다떡가	1	66	1	4	9	2	0	5	0	9	1	9	8	3	10
비빔밥	3	265	3	13	15	664	0	144	17	10	3	11	8	6	41
볶음밥	3	372	5	322	30	9	0	39	12	12	2	20	21	42	31
장국밥	0	33	0	2	1	1	0	15	1	2	0	6	4	1	11
육개장	0	17	1	1	2	1	0	10	0	5	0	6	4	3	9
이국로	4	286	0	29	21	5	2	28	2	9	3	5748	42	32	50
콩나물국	4	42	0	0	0	2	1	7	0	258	0	10	2	2	8
갈비	1	113	1	8	9	1	0	9	2	0	3	21	9	14	17
갈치	0	1	0	0	0	0	0	1	0	0	0	2	0	0	1
갈치조림	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
갈도리탕	0	33	0	1	0	2	0	4	0	3	0	15	3	2	5
고기	7	800	14	27	45	52	1	187	9	31	18	83	90	38	129
곱창	2	96	3	2	11	2	0	8	2	46	2	31	22	13	15
곰탕	1	90	2	3	13	3	0	0	0	5	5	15	17	8	10
돌어탕	1	80	1	3	4	1	0	0	6	33	6	19	13	37	12
떡볶이	7	689	14	48	56	60	4	1647	14	42	19	104	1016	70	94
김밥탕	0	79	1	2	2	3	1	10	0	0	0	3	2	78	4

Fig 5. Weight value table 2 outputted by excel

5 Conclusions

The food recommendation system is designed and implemented. The knowledge base for food recommendation in Twitter service object becomes a personalized recommendation system by extracting the related word of food and assigning weight values. So far, the recommendation system is a responsive system on request of recommendation. And the resources in this recommendation system is very limited because in specific recommendation system, the recommendations of users accessing this system are referenced.

The food recommendation system proposed in this paper, with the using of Twitter knowledge base, there are several advantages compared with traditional recommendation system. First, it is easy to collect information because SNS service such as Twitter, Facebook etc. supports the information searching API. Second, the information collection objects are varied. The people accessing SNS with varied ages, varied interesting. A large amount of information can be collected from the lots of users. Third, with the change of society the preference changes. This recommendation system covers the information of people's feeling and anniversary.

However, there is space to improve this food recommendation system. For example, more SNS services are used as the knowledge base. The implementation of recommendation system with food related vocabularies automatic analysis by using of web agents and the building of a more accurate automatic database are the future works also

6 References

[1] Wi-Geun Kim, Min-Jae Choi, "The Effect of SNS Users' Use Motivations on Using SNS and Recognizing Characteristics of SNS Messages", Korean Journal of Communication & Information, Vol. 60, pp.150-171, 2012.

[2] Hyuk-Chul Kwon, "Semantic Web and Ontology: Between Possibility and Limitation", Communications of the Korean Institute of Information Scientists and Engineer, Vol. 24 No.4., pp.11-16, 2006.

[3] Decker P. Mitra, S. Melnik, "Framework for the semantic Web: an RDF tutorial." IEEE Internet Computing, Vol. 4 Issue 6 pp.68-73, Nov .- Dec. 2000S.

[4] Smith, Wety, and McGuinness 2003, http://www.w3.org/TR/owl-features/

[5] Do-Heon Jung, Tea-Soo Kim, "A Study on the Thesaurus-based Ontology System for the Semantic Web", Korea Society for Information Management, Vol. 20 No. 3, pp.307-344, 2003.

[6] Hong Shik Yi, "On the Definition of the Food Nouns in the Dictionary", Korean Linguistics, Vol.38 No.1, pp.307-344, 2008.

ISBN: 1-60132-441-3, CSREA Press ©