

**International Journal of
Computer Science and Security
(IJCSS)**

ISSN : 1985-1553



VOLUME 4, ISSUE 2

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

Editor in Chief Dr. Haralambos Mouratidis

International Journal of Computer Science and Security (IJCSS)

Book: 2010 Volume 4, Issue 2

Publishing Date: 30-05-2010

Proceedings

ISSN (Online): 1985-1553

This work is subjected to copyright. All rights are reserved whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication of parts thereof is permitted only under the provision of the copyright law 1965, in its current version, and permission of use must always be obtained from CSC Publishers. Violations are liable to prosecution under the copyright law.

IJCSS Journal is a part of CSC Publishers

<http://www.cscjournals.org>

© IJCSS Journal

Published in Malaysia

Typesetting: Camera-ready by author, data conversion by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Table of Contents

Volume 4, Issue 2, May 2010.

Pages

- 149 - 159 Parallel Computing in Chemical Reaction Metaphor with Tuple Space
Hong Lin, Jeremy Kemp, Wilfredo Molin
- 160 - 175 A novel Data Mining algorithm for semantic web based data cloud
N.C.Mahanti, kanhaiya lal
- 176 - 182 Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK
Rizwan Ahmad, Aasia Khanum
- 183 - 198 Heuristics Based Genetic Algorithm for Scheduling Static Tasks in Homogeneous Parallel System
Kamaljit Kaur, Amit Chhabra, Gurvinder Singh
- 199 - 207 Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach
Olanrewaju Jelili Oyelade, Oladipupo, Olufunke Oyejoke

- 208 - 225 Multilevel Access Control in a MANET for a Defense Messaging System Using Elliptic Curve Cryptography
J. Nafeesa Begum, K.Kumar, V.Sumathy
- 226 - 236 FPGA Prototype of Robust Image Watermarking For JPEG 2000 With Dual Detection
Pankaj U.Lande, Sanjay N. Talbar, G.N. Shinde
- 237 - 255 Comparing The Proof By Knowledge Authentication Techniques
Stamati Gkarafli , Anastasios A. Economides
- 256 - 264 Development of Information Agent Reranking By Using Weights Measurement
Aliaa A.Youssif, Ashraf A. Darwish, Ahmed Roshdy

Parallel Computing in Chemical Reaction Metaphor with Tuple Space

Hong Lin

*Department of Computer and Mathematical Sciences
University of Houston-Downtown
1 Main Street, Houston, Texas 77002, USA*

linh@uhd.edu

Jeremy Kemp

*Department of Computer and Mathematical Sciences
University of Houston-Downtown
1 Main Street, Houston, Texas 77002, USA*

Wilfredo Molina

*Department of Computer and Mathematical Sciences
University of Houston-Downtown
1 Main Street, Houston, Texas 77002, USA*

Abstract

Methodologies have been developed to allow parallel programming in a higher level. These include the Chemical Reaction Models, Linda, and Unity. We present the Chemical Reaction Models and its implementation in IBM Tuple Space. Sample programs have been developed to demonstrate this methodology.

Keywords: Parallel Programming, Very High Level Languages, the Chemical Reaction Model, IBM Tuple Space.

1. Higher-Level Parallel Computing - Implicit Parallelism

Higher level parallel programming models express parallelism in an implicit way. Instead of imposing programmers to create multiple tasks that can run concurrently and handle their communications and synchronizations explicitly, these models allow programs to be written without assumptions of artificial sequencibility. The programs are naturally parallel. Examples of such kind of models include the Chemical Reaction Models (CRMs) [1, 2], Linda [3], and Unity [4, 5]. These models are created to address higher level programming issues such as formal program specification, program synthesis, program derivation and verification, and software architecture. Efficient implementation of these models has limited success and therefore obscures its direct applications in software design [6, 7]. Despite this limitation, efforts have been made in both academic and industrial settings to avail these models in real-world programming. For example, Unity has been used in industrial software design and found successful; execution efficiency of Linda has been affirmed by experiments and it is implemented by IBM Tuple Space. Recent discussions of these models in multi-agent system design have also been found in literature [8]. In the following discussion, we focus on the Chemical Reaction Models and its applications.

The Chemical Reaction Models describe computation as “chemical reactions”. Data (the “solution”) are represented as a multiset. A set of “reaction” rules is given to combine elements in

the multiset and produce new elements. Reactions take place until the solution becomes inert, namely there are no more elements can be combined. The results of computation are represented as the inert multiset. Gamma is a kernel language in which programs are described in terms of multiset transformations. In Gamma programming paradigm, programmers can concentrate on the logic of problem solving based on an abstract machine and are free from considering any particular execution environment. It has seeded follow-up elaborations, such as Chemical Abstract Machine (Cham) [9], higher-order Gamma [10, 11], and Structured Gamma [12]. While the original Gamma language is a first-order language, higher order extensions have been proposed to enhance the expressiveness of the language. These include higher-order Gamma, hmm-calculus, and others. The recent formalisms, γ -Calculi, of Gamma languages combine reaction rules and the multisets of data and treat reactions as first-class citizens [13-15]. Among γ -Calculi, γ_0 -Calculus is a minimal basis for the chemical paradigm; γ_c -Calculus extends γ_0 -Calculus by adding a condition term into γ -abstractions; and γ_n -Calculus extends γ_0 -Calculus by allowing abstractions to atomically capture multiple elements. Finally, γ_{cn} -Calculus combines both γ_c -Calculus and γ_n -Calculus. For notational simplicity, we use γ -Calculus to mean γ_{cn} -Calculus from this point on.

The purpose of the presented study is to investigate a method for implementing γ -Calculus using IBM Tuple Space. TSpace supports network computation in client/server style. The target of this effort is to enable higher order programming in a parallel computing platform, such as computer clusters, and allow for refinement of the executable programs using transformation techniques.

The paper will be organized as follows. In Section 2, we give a brief introduction to γ -Calculus. In Section 3, we discuss the method for implementing γ -Calculus in IBM Tuple space. Program examples are presented in Section 4. We conclude in Section 5.

2. γ -Calculus

The basic term of a Gamma program is molecules (or γ -expressions), which can be simple data or programs (γ -abstractions). The execution of the Gamma program can be seen as the evolution of a solution of molecules, which react until the solution becomes inert. Molecules are recursively defined as constants, γ -abstractions, multisets or solution of molecules. The following is their syntax:

| | | |
|-------|-------------------------------|-------------------------|
| M ::= | 0 1 ... 'a' 'b' ... | ; constants |
| | $\gamma P[C].M$ | ; γ -abstraction |
| | M_1, M_2 | ; multiset |
| | $\langle M \rangle$ | ; solution |

The multiset constructor “,” is associative and commutative (AC rule). Solutions encapsulate molecules. Molecules can move within solutions but not across solutions. γ -abstractions are elements of multisets, just like other elements. They can be applied to other elements of the same solution if a match to pattern P is found and condition C evaluates to true and therefore facilitate the chemical reaction. The pattern has the following syntax:

$$P ::= x \mid P, P \mid \langle P \rangle$$

where x is a variable. In addition, we allow for the use of tuples (written $x_1 : \dots : x_n$) and names of types. For example, γ -abstraction

$$\gamma(x: \text{Int}, y: \text{Int})[x \geq y].x$$

can be interpreted as: replace x, y by x if $x \geq y$, which is equivalent to finding the maximum of two integers.

The semantics of γ -Calculus is defined as the following:

$$\begin{array}{ll}
 (\gamma p[c].m_1), m_2 & = \phi m_1 \text{ if } \text{match}(p/m_2) = \phi \text{ and } \phi c & ; \gamma\text{-conversion} \\
 m_1, m_2 & = m_2, m_1 & ; \text{commutativity} \\
 m_1, (m_2, m_3) & = (m_1, m_2), m_3 & ; \text{associativity} \\
 E_1 = E_2 & \Rightarrow E[E_1] = E[E_2] & ; \text{chemical law}
 \end{array}$$

The γ -conversion describes the reaction mechanism. When the pattern p matches m_2 , a substitution ϕ is yielded. If the condition ϕc holds, the reactive molecules $\gamma p[c].m_1$ and m_2 are consumed and a new molecule ϕm_1 is produced. $\text{match}(p/m)$ returns the substitution corresponding to the unification of variables if the matching succeeds, otherwise it returns fail.

Chemical law formalizes the locality of reactions. $E[E_1]$ denotes the molecule obtained by replacing holes in the context $E[]$ (denoted by $[]$) by the molecule E_1 . A molecule is said to be inert if no reaction can be made within:

$$\begin{array}{l}
 \text{Inert}(m) \quad \square \\
 (m \equiv m'[(\gamma p[c].m_1), m_2] \Rightarrow \text{match}(p/m_2) = \text{fail})
 \end{array}$$

A solution is inert if all molecules within are inert and normal forms of chemical reactions are inert γ -expression. Elements inside a solution can be matched only if the solution is inert. Therefore, a pattern cannot match an active solution. This ensures that solutions cannot be decomposed before they reach their normal form and therefore permits the sequentialization of reactions. The following inference rule governs the evolution of γ -expressions:

$$\frac{E_1 \rightarrow E_2 \quad E \equiv C[E_1] \quad E' \equiv C[E_2]}{E \rightarrow E'}$$

This high level language can be implemented in Java using IBM's TSpaces server and Java package. The method is detailed in the following section.

3. IBM Tuple Space

IBM Tuple Space was originally invented as an implementation of Linda computational model. While version 3.0 is only available after obtaining a site license, version 2.12 is freely available. Installing TSpaces is as easy as unpackaging the TSpaces package on the networked file system (NFS), adding it's directory to the users classpath, and starting the server in a GNU Screen session.

3.1 Data Structures and Methods

A TSpaces program uses more of a client/server model, with each node only communicating with a 'host' or any machine with a known name, which on a cluster would usually be the head node. And although TSpaces is flexible enough to assign global names and ranks to each node, micro-managing the communications, this would defeat the purpose of having the abstraction layer TSpaces offers you. Data in a TSpaces program is shared through Tuples, a tuple is a data structure that wraps all other data structures in a tuplespace, this can include data primitives, as well as standard java classes and user defined classes. Every tuple is in a TupleSpace, and every TupleSpace has a host (actually a fully qualified host name).

The TSpaces methods used to obtain tuples are: `read()`, `waitToRead()`, `take()`, `waitToTake()`. The `read()` and `take()` methods vary in that the `read()` method leaves the tuple returned to the program in the tuplespace, and the `take()` method removes it from the tuplespace. The `waitTo` versions wait until a tuple appears in a tuplespace, and then takes it; these are good for synchronization, and can be set to time out in order to prevent the program from hanging indefinitely. These

methods take Tuples as arguments, and return the first tuple from the tuplespace that matches either the data types you specified, or specific values. There is also the scan() method, which works like read() except it returns all values that match the tuple specified. There are other Tspace methods that allow you more control over your program; countN() returns all the matching tuples and delete() removes all matching tuples. There are many other advanced features that allow you to manipulate TSpaces with a great deal of precision.

Another bonus of using a platform that runs on Java is being able to run these programs on any OS, and even inside a web browser using an applet. Such applets are ideal for being able to monitor or visualize a reaction that is running.

3.2 Synchronization

The most difficult task in writing a CRM/gamma program in Tspaces is synchronization. Synchronization is needed to determine when to terminate the program, or when the molecules stop reacting.

One program acts as one molecule, and different programs can read from the same tuplespace, but in order to test that the reaction of a molecule is exhausted, a program must run through the entire data space. So, reactions that involve more than one molecule type should be run in a cycle, the first program reacting all of the molecules from the original tuplespace, and writing the resulting tuples, as well as all of the unused tuples to the next tuplespace. The next program acts on this tuplespace until it is stable with regard to the molecule it represents, leaving the tuples in a third tuplespace. This continues until all of the programs that represent molecules have had a chance to run, and then it starts over. Termination occurs when the programs cycle through all of the programs without any changes to the tuples.

Molecules that reduce tuplespaces to a single solution are the simplest. They require only one tuplespace and no synchronization. The max number finder and the tuple adder are examples of this, they simply react until there is only one element left, then they are done.

After some further thought (but not further programming) a relation of numbers of tuplespaces to input and output elements of the gamma function can be noticed. In order to use the style of synchronization used in the sorting program, a program with X input molecules and Y output molecules requires X tuplespaces for the input and Y tuplespaces for the output. In order to detect stability, it will have to empty the input tuples 1 time with no changes and empty the output tuplespace 1 time.

The sorting program is an example of this; the method it uses involves alternating between comparing two tuples of the form (even index n, dataValue), (n + 1, dataValue) and (odd index m, dataValue), (m + 1, dataValue), and incrementing a counter tuple whenever a change occurs. This allows a comparison after the last tuples are removed from the tuplespace and into another tuplespace to determine if any changes have been made. If two consecutive changeless runs occur, then every data element is in order, and the program terminates.

There is a complication with programs where number of inputs is not equal to the number of outputs. The indexing must be handled specially or removed entirely; with numbers removed, there will be n + 1 elements missing, and with numbers added, there will be elements that need to be added somewhere, these can usually be appended to the end?

If there are more inputs than outputs, then the tuplespace will eventually be reduced to the number of outputs for the one molecule. These types of programs can use the termination style of the max element and tuple adder programs; simply running until there are Y elements left, and then stopping. The only synchronization required is when emptying a tuplespace(or set of tuplespaces), and to prevent deadlock when (number of elements) < (number of processors) * (number of inputs), but this can be handled by detecting null reads and random timeouts on the waitToRead() calls.

Although Tuple Space was initially implemented to support Linda computation model, its functions well suite in the operational semantics of the chemical reactions models. We propose implementing γ -Calculus on Tuple Space. In the following, we demonstrate a few sample programs in γ -Calculus and their Tuple Space implementation.

4. Examples

4.1 Max Number Finder

Given a set of values of an ordered type M , this program returns the maximum number of the set. The following γ -abstraction compares two randomly selected values. If the first value is greater than or equal to the second, it removes the second value from the set:

$\text{select} = \gamma(a: M, b: M)[a \geq b]. a: M, \text{select}$

No global control is imposed on the way multiset elements are selected to ignite the reaction. If select is placed in an initial set M_0 of values, it will compare two values and erase the smaller at a time till the maximum is left. So the maximum number program can then be written as:

$\text{Max } M_0 = \langle \text{select}, M_0 \rangle$

If the multiset M_0 is represented as a tuple space, this program can be converted into one that finds and displays the greatest tuple inside a tuple space. It works with each node taking two tuples from the tuple space, comparing them, and placing the greatest one back to the tuple space. This process repeats itself until the termination condition is met, that is, when there is only one tuple left in the tuple space. When a node picks tuples up, if both tuples happen to be the same size, it simply places one of them back in the tuplespace while discarding the other one. If a node happens to take only one tuple because another node already picked the last remaining ones in the tuple space, this puts it back and repeats the process. This ensures that by the next check, a node will be able to take two tuples and perform the remaining operations to find the greatest tuple. If a node sees no tuples in the tuple space, this displays a message and terminates. If a node sees only one tuple in the tuple space, it assumes the greatest tuple was already found, displays a message and terminates.

Check appendix A for an example of the code implemented in TSpaces as well as its flowchart.

4.2 Tuple Adder

Given a set of values of numerical type M , we write a program to summarize all the values. The following γ -abstraction adds two randomly selected values and put the sum back into the multiset:

$\text{add} = \gamma(a: M, b: M)[\text{true}]. a+b: M, \text{select}$

The tuple adder program can then be written as:

$\text{Sum } M_0 = \langle \text{add}, M_0 \rangle$

If M_0 is represented as a tuple space, the corresponding TSpace program will add all of the tuples in a tuple space and displays their total sum. It works with each node taking two random tuples from the tuple space, adding them up, and placing a new tuple with their total sum back in the tuplespace (the tuples themselves are discarded). This process repeats itself until there is only one tuple left in the tuplespace, which is the total sum. If there are no tuples in the tuplespace before execution, the nodes display a message and terminate.

Check appendix B for the flowchart of the code implemented in TSpaces. Code is omitted because of the page limit.

4.3 Sorter

If a list is represented by multiset $M = \{(a, i) \mid a \text{ is value and } i \text{ an index and } i\text{'s are consecutive}\}$, the following recursive γ -abstraction replaces any ill-ordered pairs by two other pairs:

$$\text{sigma} = \gamma((a, i): M, (b, j): M) [i < j \wedge a > b]. (b, i): M, (a, j): M, \text{sigma}$$

It specifies that any two selected pairs (a, i) and (b, j) that satisfy the condition, $i < j \wedge a > b$ are replaced by two other pairs (b, i) and (a, j) , and a copy of itself. If sigma is placed in an initial set M_0 of pairs, it will replace ill-ordered pairs until the entire list is sorted. So a sorting program can be defined as:

$$\text{Sort } M_0 = \langle \text{sigma}, M_0 \rangle$$

In a tuple space, a similar process will happen. The program will sort all of the tuples in a tuple space in ascending order. Each tuple has an index and a value in the following format: (index, value). When two tuples, (i, x) and (j, y) from said tuple space S are taken by a node, it first checks whether $x > y \wedge i < j$. If this happens to be true, then the following swap is performed: $(i, y), (j, x)$ they are put back in the tuple space, and the tuples are in order. This process repeats itself until no more swaps can be performed, that is, when all of the tuples in a tuple space are arranged in ascending order.

As mentioned above, multiple tuplespaces are required to synchronize this 'single pool' abstraction, in this case four tuplespaces were used. There is a primary pool, where the data is initially stored and an alternate pool where the data is written as it is being processed. Each of these pools is broken in to an even and an odd pool

Check appendix C for the flowchart of the code implemented in TSpaces, the primary feature of this programming model, is that it can utilize up to $n/2$ processing nodes, where n is the number of data items being sorted.

We have tested the above TSpace programs on a PC cluster and observed the computation in multiple nodes, and how the increase of nodes divides the number of comparisons per node, and increases speed; all of this thanks to the abstractions and portability offered by TSpaces.

We want to point out that when converting a γ -Calculus program into a TSpace program, details must be added to make a working program. However, through the above examples, we can see the conversion is straightforward in sense of the computation model. Therefore, it is technically feasible to design an automatic conversion system that can parse γ -Calculus and convert the program into a TSpace program and this is the next goal of our ongoing project.

5. Conclusion

The Chemical Reaction Models are higher level programming models that address parallelism implicitly and allows programmers to focus on the logic of problem solving instead of deal with operational details in parallel computing. IBM Tuple Space supports client/server computation based on Linda model that uses a similar concept for data structures. We discuss a method for implementing a higher order chemical reaction model, γ -Calculus, in IBM Tuple Space. We present the rules for converting γ -Calculus constructs into TSpace codes and discuss the critical techniques such as synchronizations. Our work shows that the conversion is practical. Experiments are also conducted on a computer cluster.

6. Acknowledgements

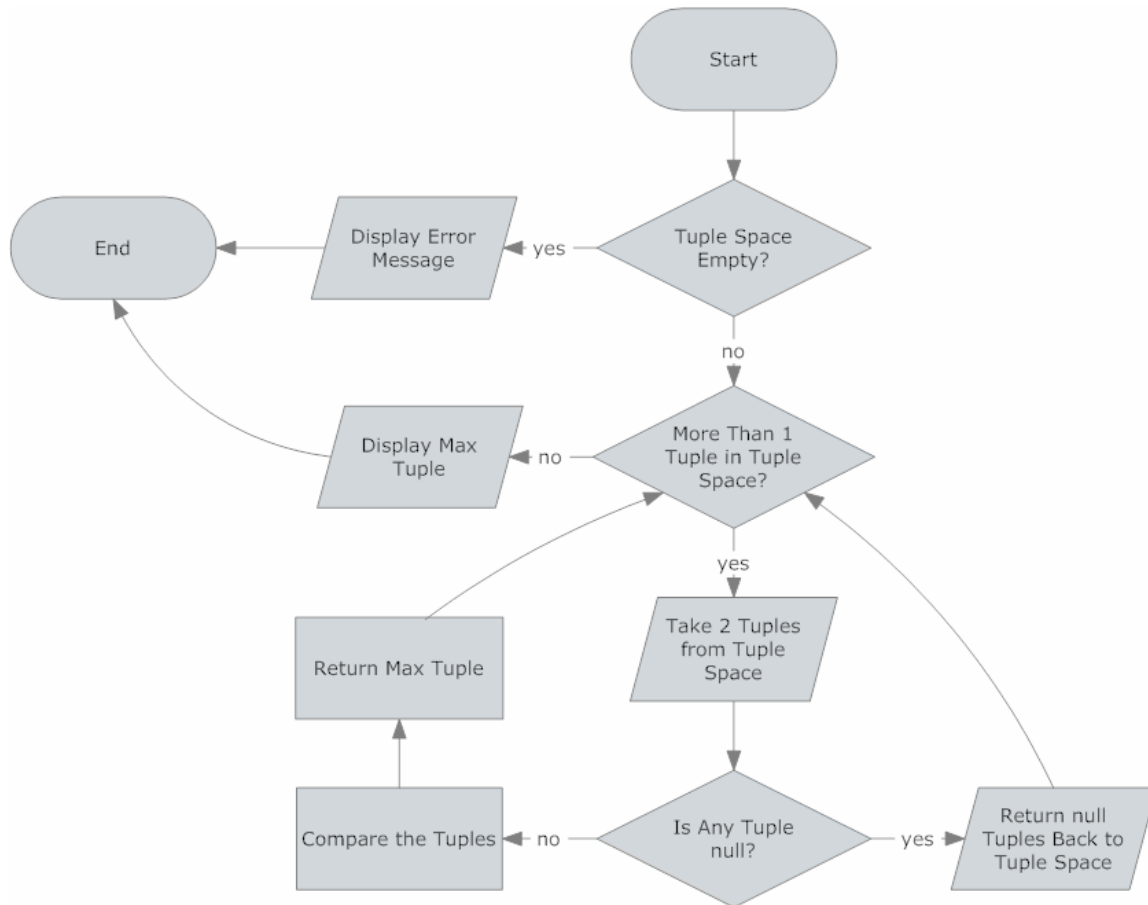
This research is partially supported by NSF grant "Acquisition of a Computational Cluster Grid for Research and Education in Science and Mathematics" (#0619312).

7. References

1. Banatre, J.-P. and Le Metayer, D. "*The Gamma model and its discipline of programming*". Science of Computer Programming, 15, 55-77, 1990.
2. Banatre, J.-P. and Le Metayer, D. "*Programming by multiset transformation*". CACM, 36(1), 98-111, 1993.
3. Carriero, N. and Gelernter, D. "*Linda in context*". CACM, 32(4), 444-458, 1989.
4. K.M. Chandy and J. Misra. "*Parallel Program Design: A Foundation*", Addison-Wesley (1988)
5. Misra, J. "*A foundation of parallel programming*". In M. Broy (ed.), Constructive Methods in Computing Science. NATO ASI Series, Vol. F55, 397-443, 1989.
6. C. Creveuil. "*Implementation of Gamma on the Connection Machine*". In Proc. Workshop on Research Directions in High-Level Parallel Programming Languages, Mont-Saint Michel, 1991, Springer-Verlag, LNCS 574, 219-230, 1991.
7. Gladitz, K. and Kuchen, H. "*Shared memory implementation of the Gamma-operation*". Journal of Symbolic Computation 21, 577-591, 1996.
8. Cabri, et al. "*Mobile-Agent Coordination Models for Internet Applications*". Computer, 2000 February, <http://dlib.computer.org/co/books/co2000/pdf/r2082.pdf>. 2000.
9. Berry, G. and Boudol, G. "*The Chemical Abstract Machine*". Theoretical Computer Science, 96, 217-248, 1992.
10. Le Metayer, D. "*Higher-order multiset processing*". DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 18, 179-200, 1994.
11. Cohen, D. and Muylaert-Filho, J. "*Introducing a calculus for higher-order multiset programming*". In Coordination Languages and Models, LNCS, 1061, 124-141, 1996.
12. Fradet, P. and Le Metayer, D. "*Structured Gamma*". Science of Computer Programming, 31(2-3), 263-289, 1998.
13. J.-P. Banâtre, P. Fradet and Y. Radenac. "*Chemical specification of autonomic systems*". In Proc. of the 13th International Conference on Intelligent and Adaptive Systems and Software Engineering (IASSE'04), July 2004.
14. J.-P. Banâtre, P. Fradet and Y. Radenac. "*Principles of chemical programming*". In S. Abdennadher and C. Ringeissen (eds.): Proc. of the 5th International Workshop on Rule-Based Programming (RULE'04), 124, ENTCS, 133-147, 2005.
15. J.-P. Banâtre, P. Fradet and Y. Radenac. "*Higher-order Chemical Programming Style*". In Proceedings of Unconventional Programming Paradigms, Springer-Verlag, LNCS, 3566, 84-98, 2005.

Appendix A – Maximum Number Finder

Flowchart



Code

```

//Wilfredo Molina - July 10, 2009
import com.ibm.tspaces.*;
public class gammaMax3
{
    public static void main(String[] args)
    {
        try
        {
            String host = "grid.uhd.edu";
            TupleSpace ts = new TupleSpace("gammaSort", host);
            Tuple template = new Tuple(new Field(Integer.class), new Field(Double.class));

            Tuple t1;
            Tuple t2;

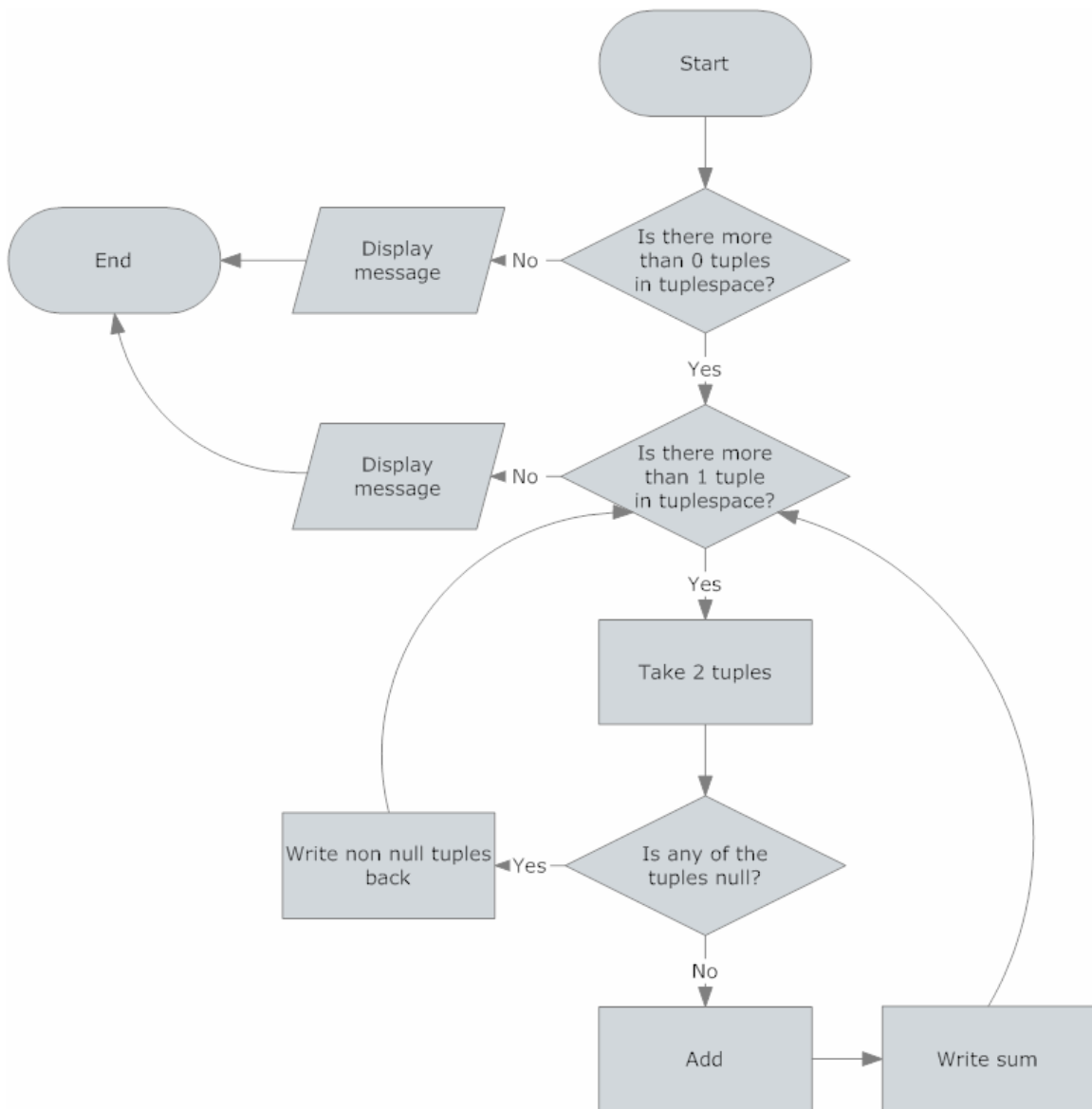
            if ((Integer)ts.countN(template) < 1)
                System.out.println("TupleSpace Empty Here");
            else
            {
                while ((Integer)ts.countN(template) > 1)
                {
                    t1 = (Tuple)ts.take(template);
                    t2 = (Tuple)ts.take(template);

                    if (t1 == null || t2 == null)
                    {
                        if (t1 != null)
                    
```

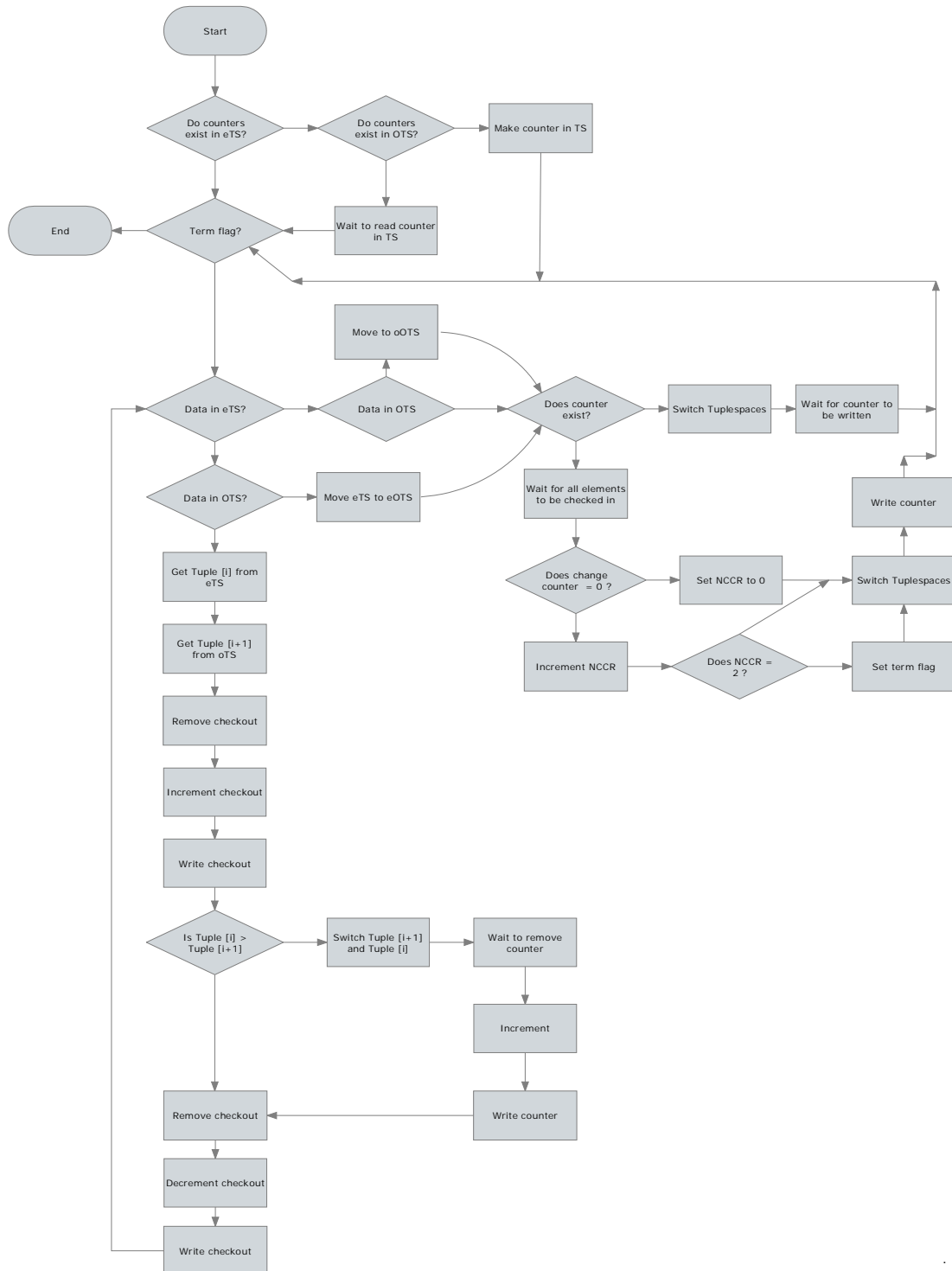
```

        ts.write(t1);
        if (t2 != null)
            ts.write(t2);
    }
    else
    {
        if ((Double)t1.getField(1).getValue() >
            (Double)t2.getField(1).getValue())
            ts.write(t1);
        else
            ts.write(t2);
    }
}
if ((Integer)ts.countN(template) == 1)
{
    t1 = (Tuple)ts.read(template);
    System.out.println("Max Found: " +
        (Double)t1.getField(1).getValue());
}
}
}
catch (TupleSpaceException tse)
{
    System.out.println("It's Broke, Wil.");
}
}
}
```

Appendix B – Tuple Adder



Appendix C – Sorter



A Novel Data Mining Algorithm for Semantic Web Based Data Cloud

Kanhaiya Lal

*Sr. Lecturer/Dept. of Computer Sc. & Engg.,
Birla Institute of Technology, Patna Campus,
Patna, 800014, India*

klal@bitmesra.ac.in

N.C.Mahanti

*Professor & Head/Department of Applied Mathematics
Birla Institute of Technology, Mesra
Ranchi, 835215, India*

ncmahanti@rediffmail.com

Abstract

By a cloud, we mean an infrastructure that provides resources and/or services over the Internet. A storage cloud provides storage services, while a compute cloud provides compute services. We describe the design of the Sector storage cloud and how it provides the storage services required by the Sphere compute cloud [14]. Different efforts have been made to address the problem of data mining in the cloud framework. In this paper we propose an algorithm to mine the data from the cloud using sector/sphere framework and association rules. We also describe the programming paradigm supported by the Sphere compute cloud and Association rules. Sector and Sphere are discussed for analyzing large data sets using computer clusters connected with wide area high performance networks

Keywords: Cloud, Web, Apriori, Sphere, Association, Sector, confidence, support.

1. INTRODUCTION

Data mining is a treatment process to extract useful and interesting knowledge from large amount of data. The knowledge modes data mining discovered have a variety of different types. The common patterns are: association mode, classification model, class model, sequence pattern and so on.

Mining association rules is one of the most important aspects in data mining. Association rules are dependency rules which predict occurrence of an item based on occurrences of other items. It is simple but effective and can help the commercial decision making like the storage layout, appending sale and etc. We usually use distributed system as a solution to mining association rules when mass data is being collected and warehoused. With the development of web and distributed techniques, we begin to store databases in distributed systems. Thus researches on the algorithm of mining association rules in distributed system are becoming more important and have a broad application foreground. Distributed algorithm has characters of high adaptability, high flexibility, low wearing performance and easy to be connected etc. [15].

The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW. The great success of the current WWW leads to a new challenge: A huge amount of data is interpretable by

humans only; machine support is limited. Berners-Lee suggests enriching the Web by machine-process able information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine process able information can point the search engine to the relevant pages and can thus improve both precision and recall. For instance, today it is almost impossible to retrieve information with a keyword search when the information is spread Fig. 1. The layers of the Semantic Web over several pages. Consider, e.g., the query for Web Mining experts in a company intranet, where the only explicit information stored are the relationships between people and the courses they attended on one hand, and between courses and the topics they cover on the other hand. In that case, the use of a rule stating that people who attended a course which was about a certain topic have knowledge about that topic might improve the results. The process of building the Semantic Web is currently an area of high activity. Its structure has to be defined, and this structure then has to be filled with life. In order to make this task feasible, one should start with the simpler tasks first. The following steps show the direction where the Semantic Web is heading:

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

Berners-Lee suggested a layer structure for the Semantic Web. This structure reflects the steps listed above. It follows the understanding that each step alone will already provide added value, so that the Semantic Web can be realized in an incremental fashion.[17]

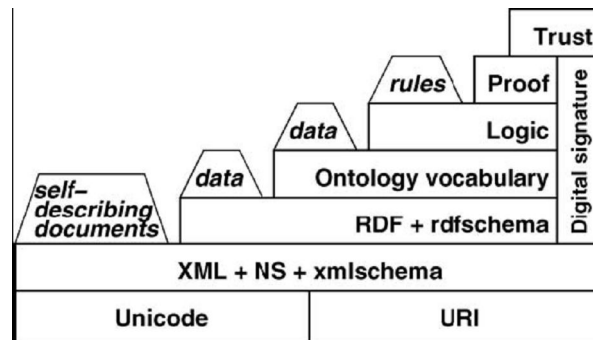


Figure 1. The layers of the Semantic Web.

The paper reveals a distributed high performance data mining system called Sector/Sphere that is based on an entirely different paradigm. Sector is designed to provide long term persistent storage to large datasets that are managed as distributed indexed files. In this paper, it has been described the design of Sector/Sphere. We also describe a data mining application developed using Sector/Sphere that searches for emergent behavior in distributed network data. Different segments of the file are scattered throughout the distributed storage managed by Sector. Sector generally replicates the data to ensure its longevity, to decrease the latency when retrieving it, and to provide opportunities for parallelism. Sector is designed to take advantage of wide area high performance networks when available [14]. The data is persistently stored and processed in place whenever possible. In this model, the data waits for the task or query. The storage clouds provided by Amazon's S3 [1], the Google File System [2], and the open source Hadoop Distributed File System (HDFS) [3] support this model. With the Sector/Sphere software from Source Forge, Terasort and Terasplit benchmarks, and the Angle datasets from the Large Data Archive, the algorithm may be implemented.

2. BACKGROUND & RELATED WORK

Cloud means, an infrastructure that provides resources and/or services over the Internet. A *storage cloud* provides storage services (block or file based services); a *data cloud* provides data management services (record-based, column-based or object-based services); and a *compute cloud* provides computational services. Often these are layered (compute services over data services over storage service) to create a stack of cloud services that serves as a computing platform for developing cloud-based applications [14].

Examples include Google's Google File System (GFS), BigTable and MapReduce infrastructure [4]; Amazon's S3 storage cloud, SimpleDB data cloud, and EC2 compute cloud [5]; and the open source Hadoop system [3]. In this section, we describe some related work in high performance and distributed data mining. For a recent survey of high performance and distributed data mining systems, see [6].

By and large, data mining systems that have been developed for clusters, distributed clusters and grids have assumed that the processors are the scarce resource, and hence shared. When processors become available, the data is moved to the processors, the computation is started, and results are computed and returned [7]. In practice with this approach, for many computations, a good portion of the time is spent transporting the data.

Key Characteristics

Agility: Agility improves with users' ability to rapidly and inexpensively re-provision technological infrastructure resources.

Cost: Cost is claimed to be greatly reduced and capital expenditure is converted to operational expenditure. This ostensibly lowers barriers to entry, as infrastructure is typically provided by a third-party and does not need to be purchased for one-time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained with usage-based options and fewer IT skills are required for implementation (in-house).

Device and location independence enable users to access systems using a web browser regardless of their location or what device they are using (e.g., PC, mobile). As infrastructure is off-site (typically provided by a third-party) and accessed via the Internet, users can connect from anywhere.

Multi-tenancy enables sharing of resources and costs across a large pool of users. One of the most compelling reasons for vendors/ISVs to utilize multi-tenancy is for the inherent data aggregation benefits. Instead of collecting data from multiple data sources, with potentially different database schemas, all data for all customers is stored in a single database schema. Thus, running queries across customers, mining data, and looking for trends is much simpler. This reason is probably overhyped as one of the core multi-tenancy requirements is the need to prevent Service Provider access to customer (tenant) information.

Centralization of infrastructure in locations with lower costs (such as real estate, electricity, etc.)

Peak-load capacity increases highest possible load-levels.

Utilization and efficiency improvements for systems that are often only 10–20% utilized.

Reliability improves through the use of multiple redundant sites, which makes cloud computing suitable for business continuity and disaster recovery. Nonetheless, many major cloud computing services have suffered outages, and IT and business managers can at times do little when they are affected.

Scalability via dynamic ("on-demand") provisioning of resources on a fine-grained, self-service basis near real-time, without users having to engineer for peak loads. Performance is monitored

and consistent and loosely-coupled architectures are constructed using web services as the system interface.

Security could improve due to centralization of data, increased security-focused resources, etc., but concerns can persist about loss of control over certain sensitive data, and the lack of security for stored kernels. Security is often as good as or better than under traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford. Providers typically log accesses, but accessing the audit logs themselves can be difficult or impossible. Furthermore, the complexity of security is greatly increased when data is distributed over a wider area and / or number of devices.

Sustainability comes through improved resource utilization, more efficient systems, and carbon neutrality. Nonetheless, computers and associated infrastructure are major consumers of energy.

Maintenance cloud computing applications are easier to maintain, since they don't have to be installed on each user's computer. They are easier to support and to improve since the changes reach the clients instantly.

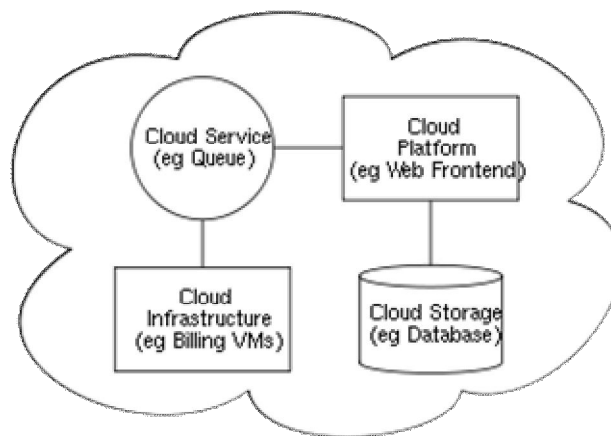


FIGURE 2: Cloud Computing Model

Cloud Computing Models

Cloud computing is a highly touted recent phenomenon. As noted, there is little hope of obtaining consensus or a standard definition regarding exactly what constitutes a “cloud” (and the term “grid” has been similarly overloaded). For example, emphasizes quality of service contracts for a cloud, contrasts social issues with technical infrastructure, while others focus on price or on the nature of the resources provided (e.g., storage, processors, platforms, or application services). Some writers emphasize what the cloud provides to its consumers, e.g., services on demand. Others emphasize what is underneath—a warehouse full of servers. The following features, especially the first three, are commonly associated with clouds. A consumer can be an individual lab, a consortium participant, or a consortium. _ Resource outsourcing: Instead of a consumer providing their own hardware, the cloud vendor assumes responsibility for hardware acquisition and maintenance. _ Utility computing: The consumer requests additional resources as needed, and similarly releases these resources when they are not needed. Different clouds offer different sorts of resources, e.g., processing, storage, management software, or application services . Large numbers of machines: Clouds are typically constructed using large numbers of inexpensive machines.

As a result, the cloud vendor can more easily add capacity and can more rapidly replace machines that fail, compared with having machines in multiple laboratories. Generally speaking

these machines are as homogeneous as possible both in terms of configuration and location. _ Automated resource management: This feature encompasses a variety of configuration tasks typically handled by a system administrator. For example, many clouds offer the option of automated backup and archival. The cloud may move data or computation to improve responsiveness. Some clouds monitor their offerings for malicious activity. _ Virtualization: Hardware resources in clouds are usually virtual; they are shared by multiple users to improve efficiency. That is, several lightly-utilized logical resources can be supported by the same physical resource. _ Parallel computing: Map/Reduce and Hadoop are frameworks for expressing and executing easily-parallelizable computations, which may use hundreds or thousands of processors in a cloud.[16]

To enable a holistic enterprise-modeling of software assets and facilitate the generalization of services across an organization and even beyond its boundaries, the Service-oriented_modeling framework (SOMF) offers virtualization capabilities:

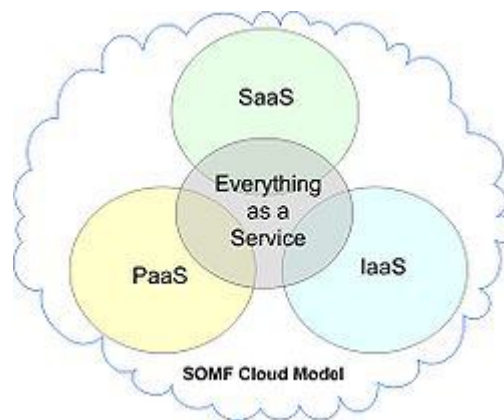


FIGURE 3: SOMF Cloud Computing Model

This ability to abstract services in spite of their location, interoperability challenges, or contribution to an architecture model fosters an elastic Cloud Computing Environment (CCE) that is nimble enough to adapt to changes and vital to business or technological imperatives.

Moreover, the Service-oriented modeling framework (SOMF) as an enterprise modeling language generalizes services. Thus, the notion of “Everything-as-a-Service” encompasses the cloud computing distributed entity model (as illustrated on the far right): infrastructure-as-a-Service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS). These can be modeled by SOMF as it conceives a service as an abstracted organizational asset and not necessarily as a Web service.

Modeling a cloud computing not only requires a language that must be able to abstract services and an environment that is typically virtual, but also hide the implementation from consumers. SOMF offers these abstraction capabilities by elevating the abstraction level of an organizational asset to enable higher cloud computing reusability rates.

Types by visibility:

Public cloud

Public cloud or external cloud describes cloud computing in the traditional mainstream sense, whereby resources are dynamically provisioned on a fine-grained, self-service basis over the Internet, via web applications/web services, from an off-site third-party provider who shares resources and bills on a fine-grained utility computing basis

Hybrid cloud

A hybrid cloud environment consisting of multiple internal and/or external providers will be typical for most enterprises. A hybrid cloud can describe configuration combining a local device, such as a Plug computer with cloud services. It can also describe configurations combining virtual and physical, collocated assets—for example, a mostly virtualized environment that requires physical servers, routers, or other hardware such as a network appliance acting as a firewall or spam filter

Private cloud

Private cloud and internal cloud are neologisms that some vendors have recently used to describe offerings that emulate cloud computing on private networks. These products claim to deliver some benefits of cloud computing without the pitfalls, capitalising on data security, corporate governance, and reliability concerns. They have been criticized on the basis that users "still have to buy, build, and manage them" and as such do not benefit from lower up-front capital costs and less hands-on management, essentially lacking the economic model that makes cloud computing such an intriguing concept.

While an analyst predicted in 2008 that private cloud networks would be the future of corporate IT, there is some uncertainty whether they are a reality even within the same firm. Analysts also claim that within five years a "huge percentage" of small and medium enterprises will get most of their computing resources from external cloud computing providers as they "will not have economies of scale to make it worth staying in the IT business" or be able to afford private clouds. Analysts have reported on Platform's view that private clouds are a stepping stone to external clouds, particularly for the financial services, and that future data centres will look like internal clouds.

The term has also been used in the logical rather than physical sense, for example in reference to platform as service offerings, though such offerings including Microsoft's Azure Services Platform are not available for on-premises deployment.

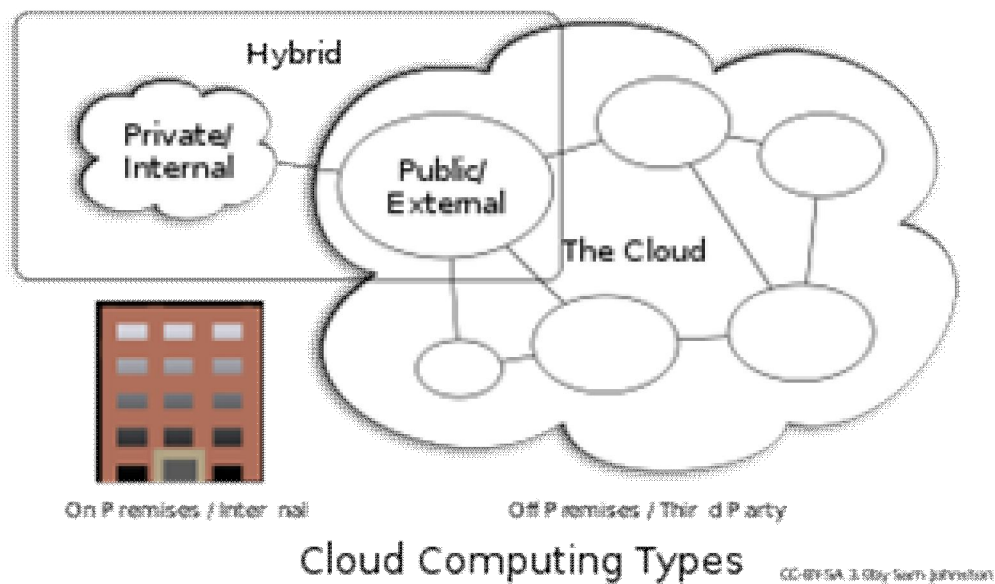


FIGURE 4: Cloud Computing Types

3. ASSOCIATION RULES.

Definition 1 confidence

Set up $I=\{i_1,i_2,i_m\}$ for items of collection, for item in $i_j(1\leq j\leq m)$, $(1\leq j\leq m)$ for lasting item, $D=\{T_1,T_N\}$ it is a trade collection, $T_i\subseteq I$ ($1\leq i\leq N$) here T is the trade. Rule $r \rightarrow q$ is probability that concentrates on including in the trade.

The association rule here is an implication of the form $r \rightarrow q$ where X is the conjunction of conditions, and Y is the type of classification. The rule $r \rightarrow q$ has to satisfy specified minimum support and minimum confidence measure. The support of Rule $r \rightarrow q$ is the measure of frequency both r and q in D $S(r) = |r|/|D|$

The confidence measure of Rule $r \rightarrow q$ is for the premise that includes r in the bargain descend, in the meantime includes q $C(r \rightarrow q) = S(r_q) / S(r)$

Definition 2 Weighting support

Designated ones project to collect $I = \{i_1, i_2, i_m\}$, each project i_j is composed with the value w_j of right ($0 \leq j \leq 1, 1 \leq j \leq m$). If the rule is $r \rightarrow q$, the weighting support is

$$S_w(r) = \frac{1}{k} \sum_{i \in r} w_j S(r)$$

And, the K is the size of the Set r_q of the project. When the right value w_j is the same as i_j , we are calculating the weighting including rule to have the same support.

Association rule mining is the current hot. In association rule mining algorithms, the most algorithms are based on Apriori algorithm to calculate, and in the mining process they can produce amount of option set, which reduce the efficiency of the association rule mining; at the same time the association rule mining will obtain amount of redundant rules, which will reduce the validity of the association rule mining; and also the user interaction performance of the association rule mining is relatively poor. On the basis of in-depth study on the existing data mining algorithms, according to the disadvantages of the association rule mining algorithms in the relational databases, a new data mining algorithm based on association rule is presented. Existing mining algorithm of association rules can be broadly divided into search algorithms, hierarchical algorithms, data sets partitioning algorithm, and so on[11].

1) Search algorithms

Search algorithm is to deal with all the term sets contained in the affairs which were read into the data set, so it needs to calculate the support of all term sets in the data aggregate D . Search algorithm can find all the frequent term sets only through one scan on data sets, an affair contained n projects will generate $2n - 1$ term sets, and when the terms the data set D contained is very large, the quantity of the option sets should be calculated and stored is often very large. Therefore, such algorithms can only be applied in the association rule mining with relatively concentrative data.

2) Hierarchy algorithm

The hierarchy algorithm with Apriori algorithm to be representation is to find frequent term sets since childhood and until now in the terms contained. Apriori algorithm will find all the frequent k term sets in the first k scan on the data sets, the option sets in the first $k + 1$ scan will be generated by all frequent k term sets through connecting computing. The number Apriori algorithm need to scan data sets is equal to the term numbers of the maximum frequent term sets. The hierarchical algorithm with Apriori algorithm to be representation can generate a relatively small option set, and the number of scanning database is decided by the term numbers of the maximum frequent term sets.

3) Partitioning algorithm

Data set partitioning algorithm includes partition algorithm, DIC algorithm will divide the entire data set into data blocks which can be stored in memory to handle, in order to save the I/O spending of visiting rendering. Partition algorithm only requires two times of scan on the entire data set. DIC algorithm can identify all the frequent term sets through the two times of scan when appropriately dividing data blocks. The number of the option term set of the data set dividing the algorithm generally larger than it of Apriori algorithm, increasing the data distortion can reduce the number of the option term set. Data set partitioning algorithm is the basis of the various parallel association rule mining algorithm and distributed association rule mining algorithm.

4) Apriori Algorithm

Apriori algorithm is an effective algorithm to mine Boolean association rule frequent term sets. Apriori algorithm uses the strategy of breadth-first search, that is, layer-by-layer search iterative method, first of all, find out the frequent term set with length of 1 which is recorded as L_1 , L_1 is used to find the aggregate L_2 of frequent 2-term sets, L_2 is used to find the aggregate L_3 of frequent 3-term sets, and so the cycle continues, until no new frequent k - term sets can be found. Finding each L_k needs a database scan. Finding all the frequent term sets is the core of association rule mining algorithm and it has the maximum calculating workload. Afterward, according to the minimum confidence threshold the effective association rules can be constructed from the frequent term sets[10].

5) Sphere

The Sphere Compute Cloud is designed to be used with the Sector Storage Cloud. Sphere is designed so that certain specialized, but commonly occurring, distributed computing operations can be done very simply. Specifically, if a user defines a function p on a distributed data set a managed by Sector, then invoking the command

```
sphere.run(a, p);
```

applies the user defined function p to each data record in the dataset a . In other words, if the dataset a contains 100, 000, 000 records $a[i]$, then the Sphere command above replaces all the code required to read and write the array $a[i]$ from disk, as well as the loop:

```
for (int i = 0, i < 100000000; ++i)  
p(a[i]);
```

The Sphere programming model is a simple example of what is commonly called a stream programming model. Although this model has been used for some time, it has recently received renewed attention due to its use by the general purpose GPU (Graphics Processing Units) community. Large data sets processed by Sphere are assumed to be broken up into several files. For example, the Sloan Digital Sky Survey dataset [12] is divided up into 64 separate files, each about 15.6 GB in size. The files are named *sdss1.dat*, . . . , *sdss64.dat*. Assume that the user has written a function called *find-BrownDwarf* that given a record in the SDSS dataset, extracts candidate Brown Dwarfs. Then to find brown dwarfs in the Sloan dataset, one uses the following Sphere code:

```
Stream sdss;
sdss.init(...); //init with 64 sdss files
Process* myproc = Sector::createJob();
myproc->run(sdss, "findBrownDwarf");
myproc->read(result);
```

With this code, Sphere uses Sector to access the required SDSS files, uses an index to extract the relevant records, and for each record invokes the user defined function *find- BrownDwarf*. Parallelism is achieved in two ways. First, the individual files can be processed in parallel. Second, Sector is typically configured to create replicas of files for archival purposes. These replicas can also be processed in parallel. An important advantage provided by a system such as Sphere is that often data can be processed in place, without moving it. In contrast, a grid system generally transfers the data to the processes prior to processing [7].

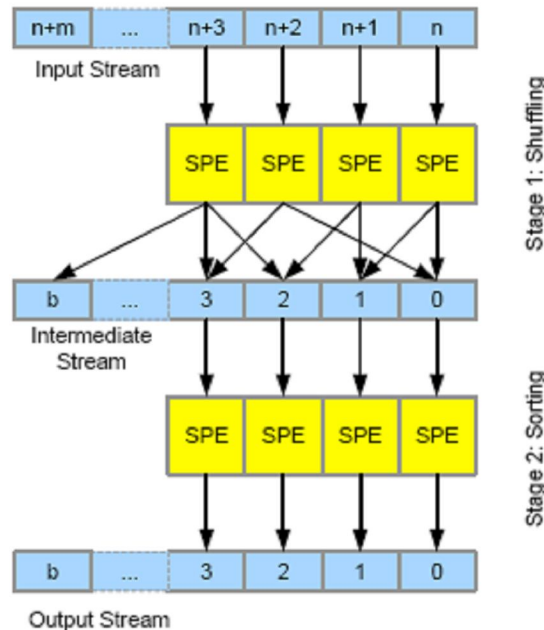


FIGURE 5: Sphere operators process Sphere streams over distributed Sphere Processing Elements (SPE) [14].

6) Sector

Sector storage cloud described as designed for wide area, high performance 10 Gb/s networks and employs specialized protocols, such as UDT to utilize the available bandwidth on these networks. Sector provides services that rely in part on the local native file systems.

The characteristics of sector are

1. Sector is designed to support a community of users, not all of whom may have write access to the Sector infrastructure.
2. Sector provides long term archival storage and access for large distributed datasets.
3. Sector is designed to utilize the bandwidth available on wide area high performance networks.
4. Sector supports a variety of different routing and network protocols. [14].

Sector has separate layers for routing and transport and interfaces with these layers through well defined APIs. In this way, it is relatively straightforward to use other routing or network protocols. In addition, UDT is designed in such a way that a variety of different network protocols can be used simply by linking in one of several different libraries. As an example, Sector is used to archive and to distribute the Sloan Digital Sky Survey (SDSS) to astronomers around the world. Using Sector, the SDSS BESTDR5 catalog, which is about 1.3TB when compressed, can be transported at approximately 8.1 Gb/s over a 10 Gb/s wide area network with only 6 commodity servers [14].

Sector assumes that large datasets are divided into multiple files, say file01.dat, file02.dat, etc. It also assumes that each file is organized into records. In order to randomly access a record in the data set, each data file in Sector has a companion index file, with a post-fix of ".idx". Continuing the example above, there would be index files file01.dat.idx, file02.dat.idx, etc. The data file and index file are always co-located on the same node. Whenever Sector replicates the data file, the index file is also replicated. The index contains the start and end positions (i.e., the offset and size) of each record in the data file. For those data files without an index, Sphere can only process them at the file level, and the user must write a function that parses the file and extracts the data

A Sector client accesses data using Sector as follows:

1. The Sector client connects to a known Sector server S, and requests the locations of an entity managed by Sector using the entity's name.
2. The Sector Server S runs a look-up inside the server network using the services from the routing layer and returns one or more locations to the client. In general, an entity managed by Sector is replicated several times within the Sector network. The routing layer can use information involving network bandwidth and latency to determine which replica location should be provided to the client.
3. The client requests a data connection to one or more servers on the returned locations using a specialized Sector library designed to provide efficient message passing between geographically distributed nodes. The Sector library used for messaging uses a specialized protocol developed for Sector called the Group Messaging Protocol.

4. All further requests and responses are performed using a specialized library for high performance network transport called UDT [8]. UDT is used over the data connection established by the message passing library.

| | | |
|-----------------------------------|-----|------------------------|
| Sector Application 1 | ... | Sector Application n |
| File Location and Access Services | | |
| Distributed Storage Services | | |
| Routing Services | | |
| Network Transport Services | | |

FIGURE 6: Sector consist of several layered services

Sector is designed to support a variety of different routing and networking protocols. The version used for the experiments described below are designed to support large distributed datasets, with loose management provided by geographically distributed clusters connected by a high performance wide area network. With this configuration, a peer-to-peer routing protocol (the Chord protocol described in [9]) is used so that nodes can be easily added and removed from the system. The next version of Sector will support specialized routing protocols designed for wide area clouds with uniform bandwidth and approximately equal RTT between clusters, as well as non-uniform clouds in which bandwidth and RTT may vary widely between different clusters of the cloud. Data transport within Sector is done using specialized network protocols. In particular, data channels within Sector use high performance network transport protocols, such as UDT [13]. UDT is a rate-based application layer network transport protocol that supports large data flows over wide area high performance networks. UDT is *fair* to several large data flows in the sense that it shares bandwidth equally between them. UDT is also *friendly* to TCP flows in the sense that it backs off when congestion occurs, enabling any TCP flows sharing the network to use the bandwidth they require. Message passing with Sector is done using a specialized network transport protocol developed for this purpose called the Group Messaging Protocol or GMP. [14].

4. METHODOLOGY

Steps for data mining using sector sphere and association rules

1. Select the minimum support threshold(T_s) and minimum confidence threshold(T_c), minimum data size($Size_{min}$) and maximum data size ($Size_{max}$).
2. We now input the data stream to the sphere processing elements. The stream is divided into data segments. The number of data segments per SPE is calculated on the basis of number of SPE and the entire stream size. Data segments from the same file are not processed at the same time until other SPE become idle.
3. The SPE accepts a new data segment from the client, which contains the file name, offset, number of rows to be processed, and additional parameters.
4. The SPE reads the data segment and its record index from local disk or from a remote disk managed by sector.
5. For each data segment find out the frequent term set with length of 1 which is recorded as L_1 , L_1 is used to find the aggregate L_2 of frequent 2-term sets, L_2 is used to find the

aggregate L_3 of frequent 3-term sets, and so the cycle continues, until no new frequent k -term sets can be found.

6. We generate strong association rules on the basis of the found frequent term sets i.e. we generate those association rules whose support and confidence respectively greater than or equal to the pre-given support threshold (T_s) and confidence threshold (T_c).
7. For each data segment (single data record, group of data records, or entire data file), the Sphere operator processes the data segment using the association rules and writes the result to a temporary buffer. In addition, the SPE periodically sends acknowledgments and feedback to the client about the progress of the processing.
8. When the data segment is completely processed, the SPE sends an acknowledgment to the client and writes the results to the appropriate destinations, as specified in the output stream. If there are no more data segments to be processed, the client closes the connection to the SPE, and the SPE is released.

Pseudocode for data mining using sector sphere & association rules

```

procedure data_mining_cloud()
{
Set minimum data segment size  $D_{min}$ 
Set maximum data segment size  $D_{max}$ 
Set minimum support threshold  $T_s$ 
Set maximum confidence threshold  $T_c$ 
Size of stream= $S$ 
Number of SPE= $N$ 
 $N_d = S/N$ ;
  for( $i=0; i < N; i++$ )
  {
     $N_{dir} = 0$ ;
  }

spawn(SPE);
for each SPE while there are data segments to be processed
{
  start a SPE connection
  If( $N_{dir} > N_d$ )
  return;
  Accept new data segment in Data[ $D_{max}$ ]
  metadata[8]=substr(Data,  $D_{max}-8$ ,  $D_{max}$ );
  findFrequentTermSets(Data,  $D_{max}-8$ );
  sendAcknowledgement( $N_{dir}$ );
   $N_{dir} = N_{dir} + 1$ ;
  release SPE connection
  return;
}
}

procedure findFrequentTermSets(Data,  $D_{max}-8$ ) // method 1//
{
  For( $k=2; L_{k-1} \neq \emptyset$ ;  $k++$ )
  {
     $C_k = \text{apriori\_gen}(L_{k-1})$ ;
    for each element  $t \in \text{Data}$ 

```

```

    {
      Ct=subset(Ck,t);
      for each candidate c∈Ct
        c.count++;
      end for;
    }
    Lk={c ∈ Ck | c.count>Ts}
  }
  return L=Uk Lk;
}

```

```

procedure apriori_gen(Lk-1){
  for each itemset I1 ∈ Lk-1
    for each itemset I2 ∈ Lk-1
      if((I1[1]=I2[1])^ (I1[2]=I2[2]) )^ (I1[3]=I2[3]) )^ (I1[4]=I2[4])^..... ]^ (I1[k-2]=I2[k-2]) ]^ (I1[k-1]=I2[k-1]) then{
        c=I1 join I2;
        if has_infrequent_subset(c,Lk-1) then
          delete c;
        else add c to Ck;
        }
      end for
    end for
  return Ck;
}

```

```

procedure has_infrequent_subset(c,Lk-1)
{
  for each k-1 subset of s of c
    If S not belongs to Lk-1 then
      return TRUE;
  return FALSE;
}

```

//Method 2//

```

For( i = 1; i < n ; i++ )
{
  Read in partition pi (pi belongs to p);
  Li=gen_i_length_itemsets(pi);
  f(x);//shifting partitions
  f(y);//shifting contents in pipeline stages to next stage
}

```

Procedure for gen i length itemsets(p_i)

```

L1p={ large 1- itemsets along with their tidlists }
For( k = 1 ; k ≤ m ; k++ )
{

```

For all itemsets I₁ ∈ L_{k-1}^p do begin

For all itemsets I₂ ∈ L_{k-1}^p do begin

If ($l_1[1]=l_2[1] \square l_1[2]=l_2[2] \square \dots \dots l_1[k-1]<l_2[k-1]$) then

```
{  
c= l1[1].l1[2] .l1[3].l1[4] .....l1[k-1].l2[k-1];  
}
```

If c cannot be pruned then

c.tidlist=l₁.tidlist \square l₂.tidlist

if(|c.tidlist|/|p|>=min. Sup.) then

$L_k^{p^*} = L_k^p \square \{ c \}$

End

End

} return $\square_k L_k^p$

For shifting partitions right (f(x))

```
{int array[n]  
for (i=n-1;i>=0;i--)  
{  
array [i]=array[i-1];  
}}
```

For shifting contents in pipeline stages to next stage (f(y))

```
{int array[m]  
for (i=m-1;i>=0;i--)  
{  
array [i]=array[i-1];  
}}
```

X={pipeline,partition}

F(pipeline);(Now shifting algorithm for pipeline)

F (partition);(shifting of partition to pipeline)

5. DISCUSSION

The efficiency of our algorithm can be highly improved if it is performed on multiprocessor system and used divide & rule method. In the Apriori algorithm, the support plays a decisive role, but now the confidence will be in the first place to mine some association rules with very high degree of confidence. The entire algorithm is divided into three steps: generate optional term sets, filter optional large term sets, and generate association rules. In the three steps, the key is filtering candidate large term sets. After a first scan on the database, the initial optional term set is generated, and after the support is calculated, the affair data term with low support is found, and then the confidence of it in the corresponding original database will be calculated through the method based on probability, if too high, the affair data term set which it belongs to will be filtered to reduce the number of records in next search; thus in the second scanning database, the scanning scope will be reduced and the searching time will be shorten, which can improve the algorithm efficiency by the same way, finally a large optional set will be generated, resulting association rules, which will be output.

6. CONSLUSION & FUTURE WORK

In this paper we described the data mining approach applied to the data cloud spread across the globe and connected to the web. We used the Sector/Sphere framework integrated to association rule based data mining. This enables the application of the association rule algorithms to the wide range of Cloud services available on the web. We have described a cloud-based infrastructure designed for data mining large distributed data sets over clusters connected with high performance wide area networks. Sector/Sphere is open source and available through Source Forge. The discovery of the association rule is a most successful and most vital duty in the data mining, is a very active research area in current data mining, its goal is to discover all frequent modes in the data set, and the current research work carrying on are mostly focused on the development of effective algorithm. On the basis of in-depth study of the existing data mining algorithms, in this paper a new data mining algorithm based on association rules is presented. The algorithm can avoid redundant rules as far as possible; the performance of the algorithm can be obviously improved when compared with the existing algorithms. We are implementing the algorithm and comparing our algorithm with other approaches.

7. REFERENCES

- [1] Amazon. Amazon Simple Storage Service (Amazon S3). www.amazon.com/s3.
- [2] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. "The Google File System". In *SOSP*, 2003.
- [3] Dhruba Borthaku. "The hadoop distributed file system: Architecture and design". retrieved from lucene.apache.org/hadoop, 2007.
- [4] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified data processing on large clusters". In *OSDI' 04: Sixth Symposium on Operating System Design and Implementation*, 2004.
- [5] Amazon Web Services LLC. "Amazon web services developer connection". retrieved from developer.amazonwebservices.com on November 1, 2007.

- [6] Hillol Kargupta. *Proceedings of Next Generation Data Mining2007*. Taylor and Francis, 2008.
- [7] Ian Foster and Carl Kesselman. “*The Grid 2: Blueprint for a New Computing infrastructure*”. Morgan Kaufmann, San Francisco, California, 2004.
- [8] Yunhong Gu and Robert L. Grossman. “UDT: UDP-based data transfer for high-speed wide area networks”. *Computer Networks*, 51(7):1777—1799, 2007.
- [9] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H Balakrishnana. “Chord: A scalable peer to peer lookup service for internet applications”. In *Proceedings of the ACM SIGCOMM '01*, pages 149–160, 2001.
- [10] Han J , Kamber M. “Data Mining: Concepts and Techniques”. 2/e San Francisco: CA. Morgan Kaufmann Publishers, an imprint of Elsevier. pp-259-261, 628-640 (2006)
- [11] Agrawal R. Imielinski T, Swami. “A Database mining: a performance perspective”. *IEEE Transactions on Knowledge and Data Engineering*, Dec.1993,5(6): 914 - 925.
- [12] Jim Gray and Alexander S. Szalay. “The world-wide telescope”. *Science*, vol 293:2037–2040, 2001.
- [13] Yunhong Gu and Robert L. Grossman. “UDT: UDP-based data transfer for high-speed wide area networks”. *Computer Networks*, 51(7):1777—1799, 2007.
- [14] Robert L. Grossman and Yunhong Gu “Data Mining using high performance data clouds: Experimental Studies using Sector and Sphere”. Retrieved from <http://sector.sourceforge.net/pub/grossman-gu-ncdm-tr-08-04.pdf>.
- [15] ZouLi & LiangXu, “Mining Association Rules in Distributed System”, First International Workshop on Education Technology and Computer Science,. IEEE, 2009.
- [16] A. Rosenthal et al. “Cloud computing: A new business paradigm for Biomedical information sharing “*Journal of Biomedical Informatics*, Elsevier ,43 (2010) 342–353 343.
- [17] G. Stumme et al. “Web Semantics: Science, Services and Agents on the World Wide Web”, *Journal of WEB Semantics*, Elsevier, 4 (2006) 124–143.
- [18] John P. Hayes ,”Computer Architecture and Organizaton”,3/e, McGraw-HILL INTERNATIONAL EDITIONS, Computer Science Series, pp- 275-292 (1998)

Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK

Rizwan Ahmad

*Computer Engineering Department
National University of Science & Technology
Rawalpindi, 46000, Pakistan*

qazirizwan.ahmad@yahoo.com

Dr. Aasia Khanum

*Computer Engineering Department
National University of Science & Technology
Rawalpindi, 46000, Pakistan*

aasia@ceme.nust.edu.pk

Abstract

Clustering is useful technique in the field of textual data mining. Cluster analysis divides objects into meaningful groups based on similarity between objects. Copious material is available from the World Wide Web (WWW) in response to any user-provided query. It becomes tedious for the user to manually extract real required information from this material. This paper proposes a scheme to effectively address this problem with the help of cluster analysis. In particular, the ROCK algorithm is studied with some modifications. ROCK generates better clusters than other clustering algorithms for data with categorical attributes. We present an enhanced version of ROCK called Enhanced ROCK (EROCK) with improved similarity measure as well as storage efficiency. Evaluation of the proposed algorithm done on standard text documents shows improved performance.

Keywords: Text Mining, Cluster Analysis, Document Similarity, Topic Generation.

1. INTRODUCTION

A considerably large portion of information present on the World Wide Web (WWW) today is in the form of unstructured or semi-structured text data bases. The WWW instantaneously delivers huge number of these documents in response to a user query. However, due to lack of structure, the users are at a loss to manage the information contained in these documents efficiently. In this context, the importance of data/text mining and knowledge discovery is increasing in different areas like: telecommunication, credit card services, sales and marketing etc [1]. Text mining is used to gather meaningful information from text and includes tasks like Text Categorization, Text Clustering, Text Analysis and Document Summarization. Text Mining examines unstructured textual information in an attempt to discover structure and implicit meanings within the text.

One main problem in this area of research is regarding organization of document data. This can be achieved by developing nomenclature or topics to identify different documents. However, assigning topics to documents in a large collection manually can prove to be an arduous task. We propose a technique to automatically cluster these documents into the related topics. Clustering is the proven technique for document grouping and categorization based on the similarity between these documents [1]. Documents within one cluster have high similarity with each another, but low similarity with documents in other clusters.

Various techniques for accurate clustering have been proposed, e.g. K-MEAN [3, 8], CURE [11, 12], BIRCH [10], ROCK [1, 2], and many others [1], [3], [10], [11]. K-MEAN clustering algorithm is used to partition objects into clusters while minimizing sum of distance between objects and their nearest center. CURE (Clustering Using Representation) represents clusters by using multiple well scattered points called representatives. A constant number 'c' of well scattered points can be chosen from '2c' scattered points for merging two clusters. CURE can detect clusters with non-spherical shapes and works well with outliers [11, 12]. BIRCH (Balance and Iterative Reducing and Clustering using

Hierarchies) is useful algorithm for data represented in vector space. It also works well with outliers like CURE [10]. However, the traditional clustering algorithms fail while dealing with categorical attributes. As they are based on distance measure so their merging processing is not accurate in case of categorical data. ROCK (Robust Clustering Algorithm for Categorical Attributes) gives better quality clusters involving categorical data as compared with other traditional algorithms. Below we first describe the original ROCK approach and then propose our own enhancements to ROCK which we call the Enhanced ROCK or EROCK approach.

2. ROCK ALGORITHM

ROCK is an agglomerative hierarchical clustering algorithm [2]. The original algorithm used the Jaccard coefficient for similarity measure but later on a new technique was introduced according to which two points are considered similar if they share a large enough number of neighbors. The basic steps of ROCK algorithm are:

1. Obtain a random sample of points from the data set 'S'
2. Compute the link value for each pair of points using the Jaccard coefficient [2]:

$$Sim (T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

3. Maintain a heap (as a sparse matrix) for each cluster's links
4. Perform an agglomerative hierarchical clustering on data using number of shared objects (as indicated by the Jaccard coefficient) as clustering criterion.
5. Assign the remaining points to the found cluster
6. Repeat steps 1-5 until the required number of clusters has been found.

ROCK algorithm has following advantages over other clustering algorithms [1]:

1. It works well for the categorical data.
2. Once a document has been added to a specific cluster, it will not be re-assigned to another cluster at the same level of hierarchy. In other words, document switching across the clusters is avoided using ROCK.
3. It uses the concept of links instead of using distance formula for measuring similarity resulting in more flexible clustering.
4. It generates better quality clusters than other algorithms.

Limitations of ROCK include the following:

1. ROCK algorithm used sparse matrix for storing cluster links.
2. Sparse matrix takes more space so efficiency suffers adversely.
3. Similarity is calculated by using Jaccard coefficient.
4. Similarity function is dependent on document length.

3. PROPOSED ENHANCEMENTS (EROCK)

EROCK approach includes several enhancements to overcome the limitations of the ROCK algorithm. Here we discuss these enhancements.

First, ROCK algorithm draws random sample from the database. It then calculates links between the points in the sample. The proposed approach (EROCK) makes use of entire data base for clustering. Every point in the database is treated as a separate cluster meaning that every document is treated as a cluster. Then the links between these clusters are calculated. The clusters with the highest number of links are then merged. This process goes on until the specified numbers of clusters are formed. So by decomposing the whole database, linkage and topic generation will become efficient.

Second, ROCK algorithm uses similarity measure based on Jaccard coefficient. We propose cosine measure:

$$CosSim (v_1, v_2) = \frac{|v_1 \cdot v_2|}{|v_1| |v_2|}$$

where v_1 and v_2 are the term frequency vectors. $v_1 \cdot v_2$ is the vector dot product defined as:

$$\sum_{i=1}^k v_{1i} v_{2i}$$

and $|v_1|$ is defined as:

$$|v_1| = \sqrt{v_1 \cdot v_1}$$

Cosine similarity is independent of the document length. Due to this property processing becomes efficient. Cosine similarity has advantages over Euclidean distance while applied on large documents (when documents tends of scale up), Euclidean will be preferred otherwise.

Third, ROCK uses sparse matrix for link information. The sparse matrix requires more space and long list of references because of which efficiency suffers adversely. In EROCK adjacency list instead of sparse matrix is proposed for maintaining link information between neighboring clusters. Adjacency list is a preferred data structure when data is large and sparse. Adjacency list keeps track of only neighboring documents and utilizes lesser space as compared to sparse matrix. Besides space efficiency it is easier to find all vertices adjacent to a given vertex in a list.

4. IMPLEMENTATION

4.1 Inputs

The EROCK algorithm requires some initial parameters which are necessary for the whole process. Following are the major inputs to run the algorithm:

- A directory containing text documents (Corpus).
- Threshold for number of clusters to be formed.
- Threshold value for measuring similarity of documents.
- Threshold value for taking top most frequent words for labeling the folders.

4.2 Document Clustering and Topic Generation Using EROCK Algorithm

Basic steps of EROCK are the same as those of ROCK. For document clustering and topic generation, the text files in the corpus are first converted into documents. Following are the steps involved in making the clusters, using EROCK algorithm:

1. Build documents from the text file present in the specified folder.
2. Compute links of every document with every other document using cosine similarity measure [2].
3. Maintain neighbors of each document in an adjacency list structure.
4. After computing links for all documents, each document is treated as a cluster.
5. Extract the best two clusters that will be merged to form one cluster. This decision is made on the basis of goodness measures. In EROCK, goodness measure defined as the two clusters which have maximum number of links between them [2]. Let these two clusters be u and v .
6. Now merge the two clusters u and v . Merging of two clusters involve, merging the names of the two clusters, the documents of two clusters and links of two clusters. This will result in a merged cluster called w .
7. For each cluster x that belongs to the link of w take following steps:
 - i. Remove clusters u and v from the links of x .

- ii. Calculate the link count for w with respect to x .
- iii. Add cluster w to the link of x .
- iv. Add cluster x to the link of w .
- v. Update cluster x in the original cluster list.
- vi. Add cluster x to the original cluster list
- vii. Repeat step (iv.) until the required number of clusters are formed or there are no two clusters found to be merged.
- viii. After obtaining the final merged cluster list apply labeling process on each. For labeling, the most frequent word from each document of a cluster is used. Take top most frequent words based on the threshold value.

The word with high frequency will be treated as the topic or label for a cluster. All related documents will be placed under one topic. Physically these documents will be put in folders with topics or label as folder name.

4.3 Output:

- A list of clusters labeled properly.
- Each cluster gets converted into a physical folder/directory on the disk and each folder contains the documents of the respective cluster.

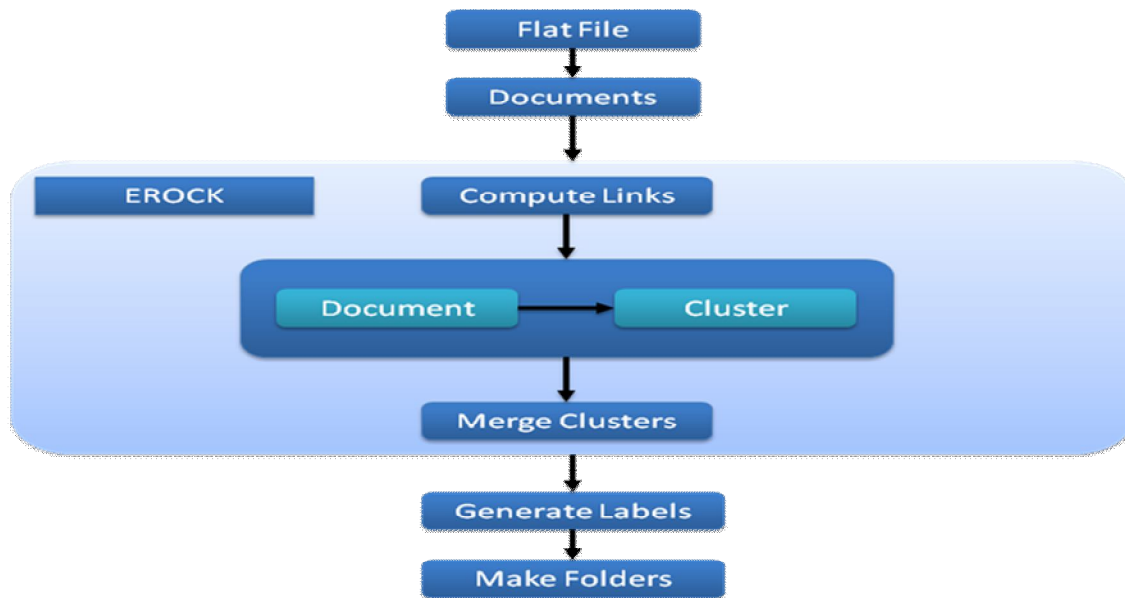


Figure 1: Application Flow

5. EXPERIMENTAL SETUP AND RESULTS

In order to evaluate the effectiveness of our algorithm, we compared the results obtained from ROCK algorithm with EROCK algorithm on similar text documents (corpus). We run both the algorithms on different corpus sizes i.e. 10, 20, 40, 50, 100, 200. For final algorithm comparison, the size of the corpus was four hundred (400) documents. Initially stop words and other useless items were removed from the documents in a pre-processing stage. The first step is to remove common words, for example, *a*, *the*, *or*, and *all* using a list of stop words. If a word in a document matches a word in the list, then the word will not be included as part of the query processing [18]. After the generation of intermediate form, clustering algorithm was applied on it. We report results from two types of analysis: Cluster Analysis and Label Analysis.

5.1 Cluster Analysis

We analyzed clustering results by varying the desired number of clusters from one (1) to ten (10). For any target number of clusters, similarity values (threshold) can have the range from 0.1 to 1.0. Figure 2 shows the overall results obtained from this study. According to the figure, clusters to be obtained are dependent on similarity threshold values. The inferences gained from the analysis are given as under:

- If the number of clusters to be obtained is equal to the number of documents then similarity factor has no effect on the clustering results.
- If the number of clusters to be obtained is less than actual number of documents, then the number of clusters to be obtained depends upon the similarity threshold.
- Increase in the threshold of top frequent word(s) of cluster will increase the size of the cluster label.
- For the dataset which we used for analysis, EROCK discovered almost pure clusters containing relevant documents with respect to their topics.

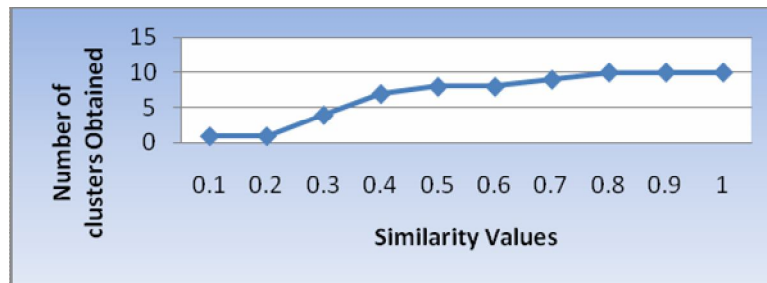


Figure 1: Clusters Obtained w.r.t Similarity values

Table 1 shows cluster labeling by similarity threshold values and number of clusters to be obtained. Here we use ten documents for summarization. If similarity value is 0.1 and number of clusters to be obtained is 1, then only one single label or topic will be generated and the entire document will be put under it. If similarity value is 0.2 and number of clusters to be obtained is 2 then two labels will be generated. If similarity value is 1.0 and numbers of clusters to be obtained are 10 then all the documents will be labeled separately as clusters. It means that labeling or document topics are mainly dependent on both similarity threshold values and number of clusters to be obtained. Other scenarios are very obvious from the table.

| Similarity Threshold | Clusters to be Obtained | Folders (Comma Separated) |
|----------------------|-------------------------|---|
| 0.1 | 1 | ALERTS ALARMS WORM AURORA COST DATA CLASSIFIERS HOST |
| 0.2 | 2 | ALERTS ALARMS WORM1 AURORA COST DATA CLASSIFIERS HOST, WORM |
| 0.3 | 3 | ALARMS, ALERTS, CLASSIFIERS, WORM AURORA COST DATA HOST |
| 0.4 | 4 | ALARMS, ALERTS, AURORA, CLASSIFIERS, DATA, HOST, WORM COST |
| 0.5 | 5 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST, DATA, HOST, WORM |
| 0.6 | 6 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST, DATA, HOST, WORM |
| 0.7 | 7 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST, DATA, HOST, WORM, WORM6 |
| 0.8 | 8 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST, COST8, DATA, HOST, WORM, |
| 0.9 | 9 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST,COST8 DATA, HOST, WORM, WORM7 |
| 1.0 | 10 | ALARMS, ALERTS, AURORA, CLASSIFIERS, COST, COST8, DATA, HOST, WORM, WORM7 |

Table 1: Cluster analysis results with topics

5.2 Labeling Analysis

Labeling analysis involve the analysis of labels generated for clusters based on similarity threshold values. This analysis is helpful to check whether the process is accurate or not. Label generation varies as per similarity vales as shown in Table 1. Following parameters were used for this analysis:

- Number of Documents: 10

- Number of Cluster to be obtained: 4
- Similarity Threshold: 0.3

| Top Label Frequency Threshold | Clusters to be Obtained | Labels (Comma Separated) |
|-------------------------------|-------------------------|---|
| 0.3 | 4 | ALARMS, ALERTS, CLASSIFIERS, WORM AURORA |
| 0.5 | 4 | ALARMS, ALERTS, CLASSIFIERS, WORM AURORA COST |
| 0.7 | 4 | ALARMS, ALERTS, CLASSIFIERS, WORM AURORA COST DATA |
| 1.0 | 4 | ALARMS, ALERTS, CLASSIFIERS, WORM AURORA COST DATA HOST |

Table 2: Cluster Labeling w.r.t Similarity Value

5.3 Comparison of ROCK & EROCK Performance

It is noteworthy that comparison should be performed on same machine and under same environment. We used the same machine for comparisons of ROCK & EROCK. It is also necessary that algorithm should also be implemented in same technology and on same platform. For this we implemented ROCK & EROCK algorithm on same technology and platform.

Performance results of both algorithms are shown in Figure 2 & Figure 3 and with similarity threshold of 0.1 and 0.5 respectively. We compared the both algorithm with varying sizes of the document (we calculated the number of words in a document, after removal of stop words). In Figure 2 & Figure 3 we mentioned the document size (No.of Words) horizontally and time (in second) vertically. It is very clear from both the figures that when document size goes on increasing, EROCK give good performance as compared with ROCK algorithm.

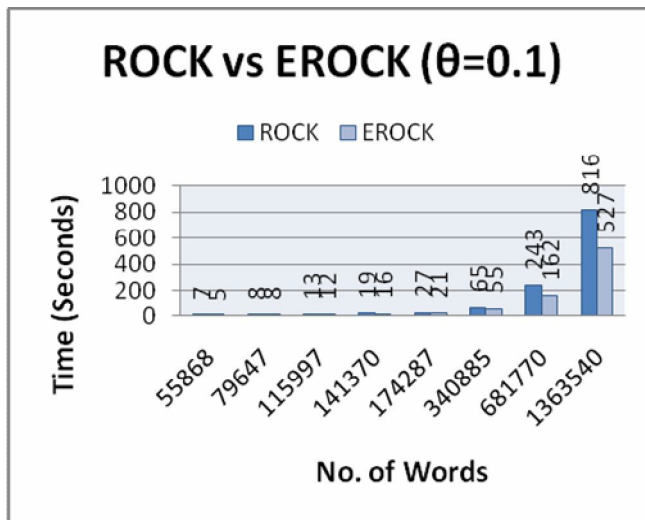


Figure 2: ROCK vs EROCK with SM=0.1

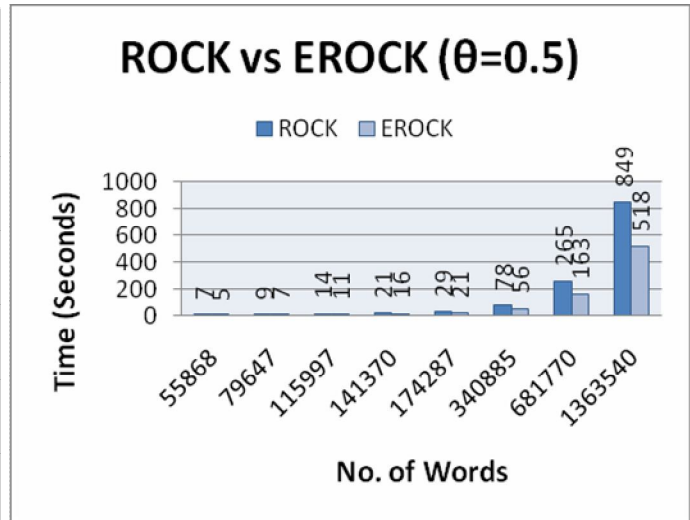


Figure 3: ROCK vs EROCK with SM=0.5

6. CONCLUSION & FUTURE WORK

In this paper we proposed an efficient way of document topic generation with enhanced version of cosine based similarity between the pair of categorical data known as clusters. We also proposed and used efficient document storage technique i.e. adjacency list instead of sparse matrix. By enhancing these two parameters of traditional ROCK algorithm, we get better results (as shown in Figure 2 & Figure 3). The experimental results obtained from the research are very encouraging. The outcome of this research shows that by using proposed approach, the cumbersome task of manually grouping and arranging files becomes very easy. Now user will be able to get relevant information easily without doing tedious manual activity. Huge information is now available in the form of text documents so documents/clusters having related information are grouped together and labeled accordingly. Clusters are merged only if closeness and inter connectivity of items within both clusters are of high significance. Finally it is observed that EROCK gives good performance for large datasets.

There are many areas in text mining; where one may carry on his/her work to enhance those areas. Out of these, the labeling of the clusters is a very daunting challenge of this time. No remarkable effort has been made in this regard to get good result. That is why automatic labeling of the clusters is not so much accurate. A keen and concerted work has been done to remove this hurdle. It will certainly serve as a lime length for future researchers.

7. REFERENCES

- [1] Shaoxu Song and Chunping Li, "Improved ROCK for Text Clustering Using Asymmetric Proximity", SOFSEM 2006, LNCS 3831, pp. 501–510, 2006.
- [2] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim, "ROCK: A robust clustering algorithm for categorical attributes". In: IEEE Internat. Conf. Data Engineering, Sydney, March 1999.
- [3] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.
- [4] Alain Lelu, Martine Cadot, Pascal Cuxac, "Document stream clustering: experimenting an incremental algorithm and AR-based tools for highlighting dynamic trends.", International Workshop on Webometrics, Informatics and Scientometrics & Seventh COLIENT Meeting, France, 2006.
- [5] Jiyeon Choo, Rachsuda Jiamthapthaksin, Chun-sheng Chen, Oner Ulvi Celepcikay, Christian Giusti, and Christoph F. Eick, "MOSAIC: A proximity graph approach for agglomerative clustering," Proceedings 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), Regensburg Germany, September 2007.
- [6] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, Proceedings of the 16th IEE International Conference on Tools with AI, 2004, pp. 576–584.
- [7] Murtagh, F., "A Survey of Recent Advances in Hierarchical Clustering Algorithms", The Computer Journal, 1983.
- [8] Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, 2, p. 283-304.
- [9] Huidong Jin , Man-Leung Wong , K. -S. Leung, "Scalable Model-Based Clustering for Large Databases Based on Data Summarization", IEEE Transactions on Pattern Analysis and Machine Intelligence, v.27 n.11, p.1710-1719, November 2005.
- [10] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases".
- [11] Linas Baltrunas, Juozas Gordevicius, "Implementation of CURE Clustering Algorithm", February 1, 2005.
- [12] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "CURE: An Efficient Clustering Algorithm for Large Databases".
- [13] M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone, "A Web Text Mining Flexible Architecture", World Academy of Science, Engineering and Technology 32 2007.
- [14] Brigitte Mathiak and Silke Eckstein, "Five Steps to Text Mining in Biomedical Literature", Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics
- [15] Ng, R.T. and Han, J. 1994. Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th VLDB Conference, Santiago, Chile, pp. 144–155.
- [16] Stan Salvador and Philip Chan, Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms, Proc. 16th IEEE Intl. Conf. on Tools with AI, pp. 576–584, 2004.
- [17] Sholom Weiss, Brian White, Chid Apte, "Lightweight Document Clustering", IBM Research Report RC-21684.
- [18] Masrah Azrifah Azmi Murad, Trevor Martin, "Similarity-Based Estimation for Document Summarization using Fuzzy Sets", IJCSS, Volume (1): Issue (4), pp 1-12.
- [19] Sohil Dineshkumar Pandya, Paresh V Virparia, "Testing Various Similarity Metrics and their Permutations with Clustering Approach in Context Free Data Cleaning", IJCSS, Volume (3): Issue (5), pp 344-350.

Heuristics Based Genetic Algorithm for Scheduling Static Tasks in Homogeneous Parallel System

Kamaljit Kaur

*Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar- 143001, Punjab, India*

kamal.aujla86@gmail.com

Amit Chhabra

*Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar- 143001, Punjab, India*

chhabra_amit78@yahoo.com

Gurvinder Singh

*Department of Computer Science & Engineering,
Guru Nanak Dev University,
Amritsar- 143001, Punjab, India*

gsbawa71@yahoo.com

Abstract

Multiprocessor task scheduling is an important and computationally difficult problem. Multiprocessors have emerged as a powerful computing means for running real-time applications, especially that a uni-processor system would not be sufficient enough to execute all the tasks. That computing environment requires an efficient algorithm to determine when and on which processor a given task should execute. A task can be partitioned into a group of subtasks and represented as a DAG (Directed Acyclic Graph), that problem can be stated as finding a schedule for a DAG to be executed in a parallel multiprocessor system. The problem of mapping meta-tasks to a machine is shown to be NP-complete. The NP-complete problem can be solved only using heuristic approach. The execution time requirements of the applications' tasks are assumed to be stochastic. In multiprocessor scheduling problem, a given program is to be scheduled in a given multiprocessor system such that the program's execution time should be minimized. The last job must be completed as early as possible. Genetic algorithm (GA) is one of the widely used techniques for constrained optimization. Performance of genetic algorithm can be improved with the introduction of some knowledge about the scheduling problem represented by the use of heuristics. In this paper the problem of same execution time or completion time and same precedence in the homogeneous parallel system is resolved by using concept of Bottom-level (b-level) or Top-level (t-level). This combined approach named as heuristics based genetic algorithm (HGA) based on MET (Minimum execution time)/Min-Min heuristics and b-level or t-level precedence resolution and is compared with a pure genetic algorithm, min-min heuristic, MET heuristic and First Come First Serve (FCFS) approach. Results of the experiments show that the heuristics based genetic algorithm produces much

better results in terms of quality of solutions.

Keywords: DAG, multiprocessor scheduling, genetic algorithm, heuristics.

1. INTRODUCTION

The problem of scheduling parallel tasks onto multiprocessors is to simply apportion a set of tasks to processors such that the optimal usage of processors and accepted computation time for scheduling algorithm are obtained [1,2]. The assumption of this paper is based on the deterministic model, that is, the number of processors, the execution time of tasks, the relationship among tasks and precedence constraints are known in advance. The precedence constraints between tasks are represented by a Directed Acyclic Graph (DAG). In addition, the communication cost between two tasks is considered to be non-negligible and the multiprocessor system is not diverse and non-preemptive, that is, the processors are homogeneous, and each processor completes the current task before the new one starts its execution.

The complexity of the scheduling problem is very depended to the DAG, the number of processors, the execution time of tasks and also the performance criteria which would to be optimized.

To date, many heuristic methods have been presented to schedule tasks on multiprocessor systems [5, 9, 10, 11, 16, 18, 19]. Also, there are many studies have been used for task scheduling based on GA [7, 8, 12, 13, 14, 15, 16, 17, 20, 21, 22, 23]. GA is a problem solving strategy, based on Darwinian evolution, which has been successfully used for optimization problems [3, 4].

The aim of this paper is to present a GA to decrease the computation time for finding a suboptimal schedule.

This paper is divided as follows: In section 2 an overview of the problem is given along with brief description of the solution methodology. Section 3 provides a more detailed heuristics based genetic algorithm. Experimental results and performance analysis are provided in section 4 and conclusion follow in section 5.

2. PROBLEM STATEMENT

In this section, more prescribed multiprocessor scheduling problem and the principles of genetic algorithms are discussed.

2.1 Task Scheduling Problem

Parallel Multiprocessor system scheduling can be classified into many different classes based on the characteristics of the tasks to be scheduled, the multiprocessor system and the availability of information. This paper focus on a deterministic scheduling problem. A deterministic scheduling problem [1, 2] is one in which all information about the tasks and the relation to each other such as execution time and precedence relation are known to the scheduling algorithm in advance. The tasks should be non-preemptive i.e. task execution must be completely done before another task takes control of the processor, and the processor environment is homogeneous. Homogeneous of processor means that the processors have same speeds or processing capabilities.

The main objective is to minimize the total task completion time (execution time + waiting time or idle time).

The multiprocessor computing consists of a set of m homogeneous processor

$$P = \{p_i: i = 1, 2, 3 \dots m\}$$

They are fully connected with each other via identical links. Figure 1 shows a fully connected three parallel system with identical link.

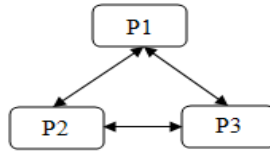


FIGURE 1: A fully connected parallel processor

Consider a directed acyclic task graph $G = \{V, E\}$ of n nodes. Each node $V = \{T_1, T_2, \dots, T_n\}$ in the graph represents a task. Aim is to map every task to a set $P = \{P_1, P_2, \dots, P_m\}$ of m processors. Each task T_i has a weight W_i associated with it, which is the amount of time the task takes to execute on any one of the m homogeneous processors. Each directed edge e_{ij} indicates dependence between the two tasks T_i and T_j that it connects. If there is a path from node T_i to node T_j in the graph G , then T_i is the predecessor of T_j and T_j is the successor of T_i . The successor task cannot be executed before all its predecessors have been executed and their results are available at the processor at which the successor is scheduled to execute. A task is "ready" to execute on a processor if all of its predecessors have completed execution and their results are available at the processor on which the task is scheduled to execute. If the next task to be executed on a processor is not yet ready, the processor remains idle until the task is ready. The elements set C are the weights of the edges as $C = \{c_k: k = 1, 2, 3 \dots r\}$ It represents the data communication between the two tasks, if they are scheduled to different processors. But if both tasks are scheduled to the same processor, then the weight associated to the edge becomes null [7, 12].

A DAG which has eleven tasks according to their height and their execution time (the time needed for a task to execute) is shown in Figure 2.

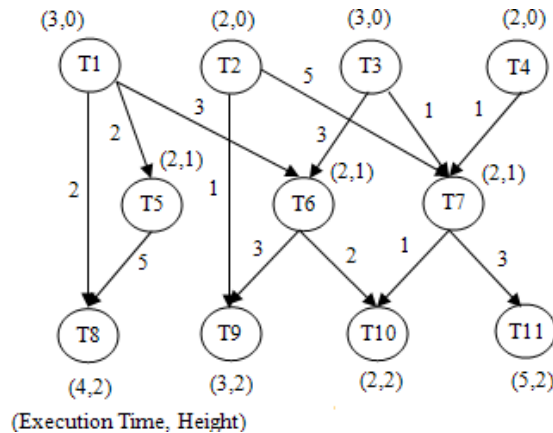


FIGURE 2: An example of a DAG

$Tlevel(T_i)$ is defined to be the length of the longest path in the task graph from an entry task to T_i , excluding the execution cost of T_i . Symmetrically, $blevel(T_i)$ is the length of the longest path from T_i to an exit task, including the execution cost of T_i . Formula (2.1) and (2.2) are formal definitions of $tlevel(T_i)$ and $blevel(T_i)$. Notice that we consider communication costs while calculating values $tlevel$ and $blevel$ [8].

$$tlevel(T_i) = \max_{T_j \in \text{pred}(T_i)} \{tlevel(T_j) + W_j + c_{ji}\} \quad (2.1)$$

$$blevel(T_i) = W_i + \max_{T_j \in \text{succ}(T_i)} \{c_{ij} + blevel(T_j)\} \quad (2.2)$$

2.2 Minimum Execution Time (MET)

Minimum Execution Time (MET) assigns each task, in arbitrary order, to the machine with the best expected execution time for that task, regardless of that machine's availability. The motivation behind MET is to give each task to its best machine [5, 11].

2.3 Min-min Heuristic

Min-min heuristic uses minimum completion time (MCT) as a metric, meaning that the task which can be completed the earliest is given priority. This heuristic begins with the set U of all unmapped tasks. Then the set of minimum completion times, $M = \{\min(\text{completion_time}(T_i, M_j)) \text{ for } (1 \leq i \leq n, 1 \leq j \leq m)\}$, is found. M consists of one entry for each unmapped task. Next, the task with the overall minimum completion time from M is selected and assigned to the corresponding machine and the workload of the selected machine will be updated. And finally the newly mapped task is removed from U and the process repeats until all tasks are mapped (i.e. U is empty) [5, 11].

2.4 Genetic Algorithms

A genetic algorithm starts with an initial population that evolves through generations and to reproduce depends on its fitness [3, 4]. In this case, the fitness of an individual is defined as the difference between its makespan and the one of the individuals having the largest makespan in the population. The best individual corresponds to the one having the smallest makespan and the largest fitness.

Next, the operators that compose a genetic algorithm are reviewed. The selection operator allows the algorithm to take biased decisions favor good individuals when changing generations. For this, some of the good individuals are replicated, while some of the bad individuals are removed. As a consequence, after the selection, the population is likely to be dominated by good individuals. Starting from a population P_1 , this transformation is implemented iteratively by generating a new population P_2 of the same size as P_1 .

Genetic algorithms are based on the principles that crossing two individuals can result an offsprings that are better than both parents and slight mutation of an individual can also generate a better individual. The crossover takes two individuals of a population as input and generates two new individuals, by crossing the parents' characteristics. The offsprings keep some of the characteristics of the parents.

The mutation randomly transforms an individual that was also randomly chosen. It is important to notice that the size of the different populations is same.

The structure of the algorithm is a loop composed of a selection followed by a sequence of crossovers and a sequence of mutations. After the crossovers, each individual of the new population is mutated with some (low) probability. This probability is fixed at the beginning of the execution and remains constant. The termination condition may be the number of iterations, execution time, results stability, etc [3, 7, 8, 6].

3. HGA: THE SUGGESTED ALGORITHM

GAs operates through a simple cycle of stages: creation of population strings, evaluation of each string, selection of the best strings and reproduction to create a new population. The number of genes and their values in each chromosome depend on the problem specification. In this paper,

the number of genes of each chromosome is equal to the number of the nodes (tasks) in the DAG and the gene values demonstrate the scheduling priority of the related task to the node (each chromosome shows a scheduling), where the higher priority means that task must be executed early. In the basic genetic algorithm the initial population is generated randomly, which can cause to generate more bad results. To avoid the generation of non-optimal results, heuristic approach along with precedence resolution can be applied to generate the initial population that gives better results in terms of quality of solutions.

3.1 Coding of Solutions

For multiprocessor scheduling problem, a schedule is one that satisfies following conditions.

1. The precedence relations among the tasks are satisfied
2. Every task is present and appears only once in the schedule [7, 8].

A schedule can be represented as several lists of computational tasks (fig 3). Each list corresponds to computational tasks executed on a processor and order of tasks in the list indicates the order of execution.

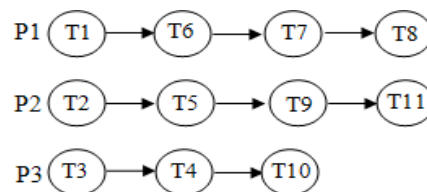


FIGURE 3: List Representation of a schedule

3.2 Population Initialization

The next step in the GAs is the creation of the initial population. Number of processors, number of tasks and population size are needed to generate initial population. Each individual of the initial population is generated through a minimum execution time or min-min heuristic along with b-level or t-level precedence resolution to avoid the problem of same execution time or completion time and same precedence. The problem of same execution time/completion time and precedence can occur in the homogeneous parallel system as all the processors take same execution time to execute one task.

The task to be scheduled for each iteration is determined by the following rules:

- i. Sort the tasks according to their execution time/completion time in ascending order according to the minimum execution time (MET)/Min-Min heuristic.
- ii. Calculate the bottom-level of each task.
- iii. Sort the tasks with the same execution time/completion time and same precedence according to their bottom-level in descending order.
- iv. Assign the tasks to the processors in the order of their bottom-level.

OR

The task to be scheduled for each iteration is determined by the following rules:

- i. Sort the tasks according to their execution time/completion time in ascending order according to the minimum execution time (MET)/Min-Min heuristic.
- ii. Calculate the top-level of each task.
- iii. Sort the tasks with the same execution time/completion time and same precedence according to their top-level in ascending order.
- iv. Assign the tasks to the processors in the order of their top-level.

The length of all individuals in an initial population is equal to the number of tasks in the DAG.

For example: the initial population of fig. 2 is generated as:

| Task Number | Execution Time | Completion Time | Bottom Level | Top Level | Order of execution According to Execution Time | Order of execution According to Completion Time | Order of execution According to Bottom-level | Order of execution According to Top-level |
|-------------|----------------|-----------------|--------------|-----------|--|---|--|---|
| 1 | 3 | 3 | 16 | 0 | 7 | 2 | 2 | 1 |
| 2 | 2 | 2 | 17 | 0 | 1 | 1 | 1 | 2 |
| 3 | 3 | 3 | 14 | 0 | 8 | 3 | 3 | 3 |
| 4 | 2 | 4 | 13 | 0 | 2 | 4 | 4 | 4 |
| 5 | 2 | 5 | 11 | 5 | 3 | 5 | 5 | 5 |
| 6 | 2 | 7 | 8 | 6 | 4 | 7 | 7 | 6 |
| 7 | 2 | 7 | 10 | 7 | 5 | 6 | 6 | 7 |
| 8 | 4 | 9 | 4 | 12 | 10 | 8 | 9 | 10 |
| 9 | 3 | 12 | 3 | 11 | 9 | 9 | 10 | 9 |
| 10 | 2 | 14 | 2 | 10 | 6 | 11 | 11 | 8 |
| 11 | 5 | 12 | 5 | 12 | 11 | 10 | 8 | 11 |

TABLE 1: Priority of execution of tasks based on their execution time, completion time, bottom-level and top-level.

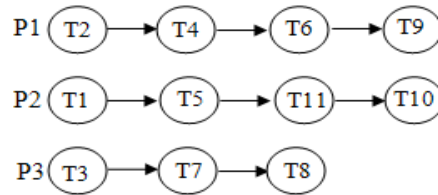


FIGURE 4: Initial Population of figure 1 using b-level resolution

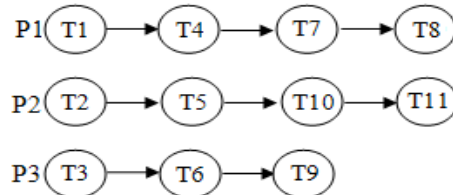


FIGURE 5: Initial Population of figure 1 using t-level resolution

3.3 Fitness Value

Several optimization criteria can be considered for this problem, certainly the problem is multiobjective in its general formulation [20]. The elementary criterion is that of minimizing the *makespan*, that is, the time when finishes the latest job. A secondary criterion is to minimize the *flowtime* that is, minimizing the sum of finalization times of all the jobs. These two criteria are defined as follows:

$$\text{makespan} : \min_{S_i \in \text{Sched}} \{ \max_{j \in \text{Jobs}} F_j \} \quad \text{and}$$

$$\text{flowtime} : \min_{S_i \in \text{Sched}} \{ \sum_{j \in \text{Jobs}} F_j \}$$

F_j denotes the time when job j finalizes, $Sched$ is the set of all possible schedules and jobs is the set of all jobs to be scheduled.

For example fitness value of the initial population is as follows:

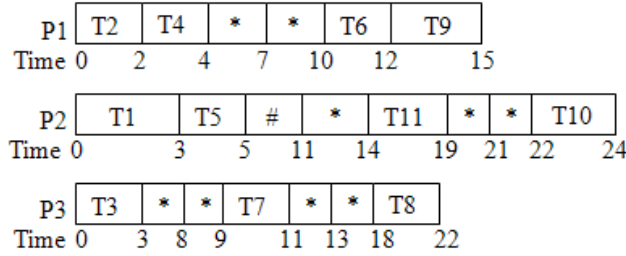


FIGURE 6: Assignment of tasks to processors using b-level resolution

The fitness value is calculated in terms of Makespan and Flowtime as discussed above as

Makespan = 24 time units

Flowtime = $3+2+3+4+5+12+11+22+15+24+19 = 120$ time units.

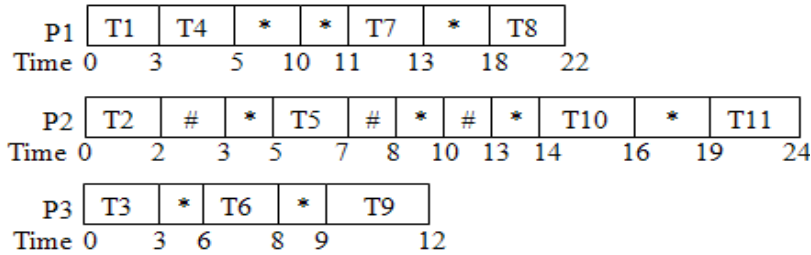


FIGURE 7: Assignment of tasks to processors using t-level resolution

Makespan = 24 time units

Flowtime = $3+2+3+5+7+8+13+22+12+16+24 = 115$ time units.

The * denotes the communication time and # denotes the waiting time.

3.4 Selection Operator

The design of the fitness function is the basic of selection operation, the design of the fitness function will directly affect the performance of genetic algorithm. GAs uses selection operator to select the superior and eliminate the inferior. The individuals are selected according to their fitness value. Once fitness values have been evaluated for all chromosomes, good chromosomes can be selected through rotating roulette wheel strategy. This operator generate next generation by selecting best chromosomes from parents and offspring.

3.5 Crossover Operator

Crossover operator randomly selects two parent chromosomes (chromosomes with higher values have more chance to be selected) and randomly chooses their crossover points, and mates them to produce two child (offspring) chromosomes. In this paper one and two point crossover operators are examined. In one point crossover, the segments to the right of the crossover points are exchanged to form two offspring as shown in figure 8 (a) and in two point crossover [3] [8], the middle portions of the crossover points are exchanged to form two offspring as shown in figure 8 (b).

Randomly selects Parent 1 & 2, crossover point 2

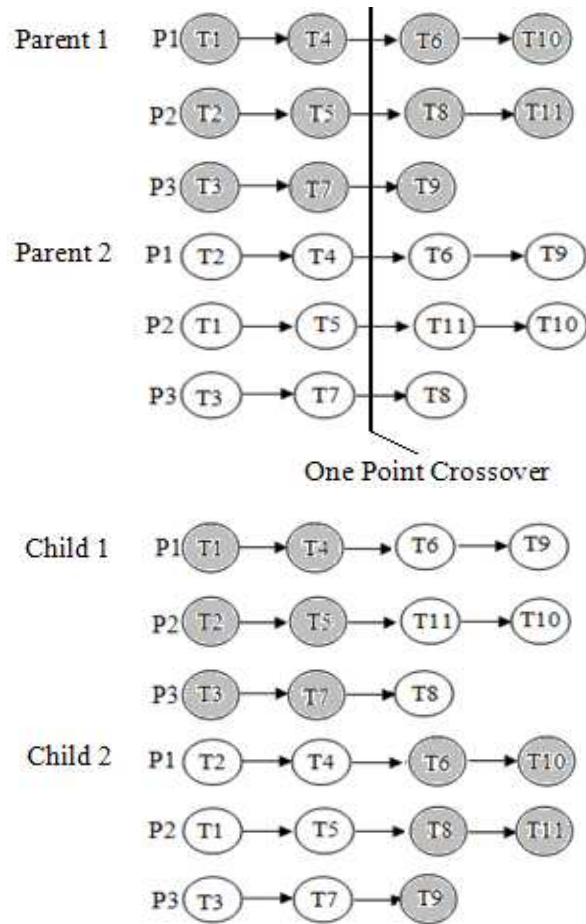
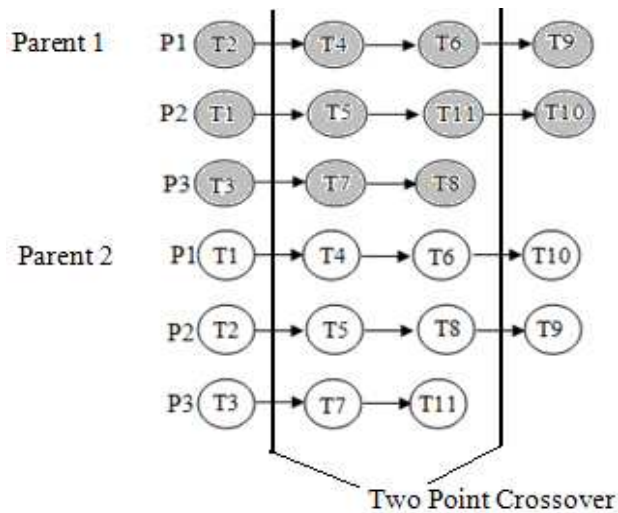


FIGURE 8(a): One Point Crossover

Randomly selects parent 1 & 2, crossover points 1 & 3



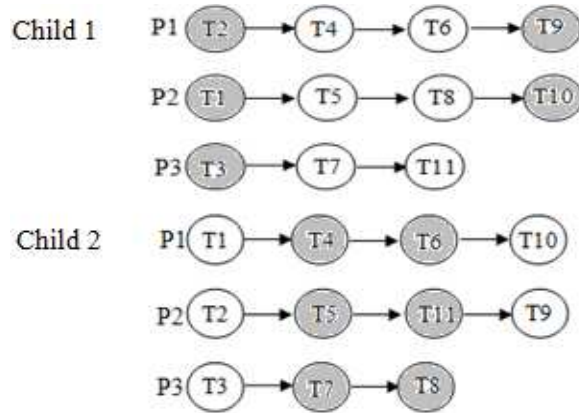


FIGURE 8(b): Two Point Crossover

3.6 Mutation

Mutation ensures that the probability of finding the optimal solution is never zero. It also acts as a safety net to recover good genetic material that may be lost through selection and crossover. Implementation of two mutation operators is there in HGA. The first one selects two tasks randomly and swaps their allocation parts. The second one selects a task and alters its allocation part at random. These operators can always generate feasible offspring, too. Figure 9(a), 9(b), 9(c) & 9(d) demonstrate the mutation operation.

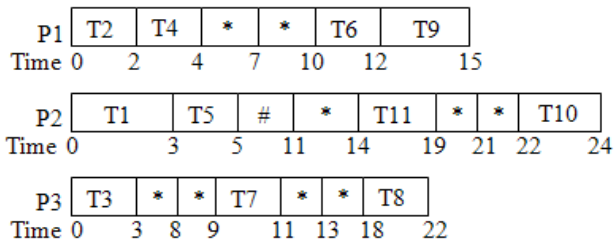


FIGURE 9(a): A Gantt chart before mutation operation

Makespan = 24 time units

Flowtime = 3+2+3+4+5+12+11+22+15+24+19 = 120 time units.

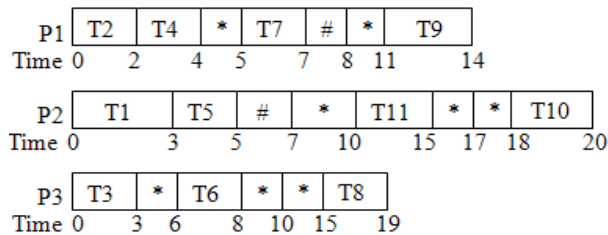


FIGURE 9(b): A Gantt chart after swap mutation operation.

Makespan = 20 time units

Flowtime = 3+2+3+4+5+8+7+19+14+20+15 = 100 time units.

The mutation operation swaps task t6 on processor p1 to task t7 on processor p3.

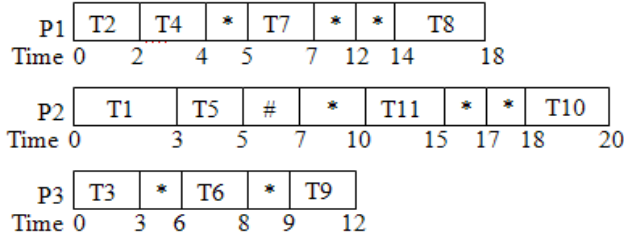


FIGURE 9(c): A Gantt chart after swap mutation operation.

Makespan = 20 time units

Flowtime = 3+2+3+4+5+8+7+18+12+20+15 = 97 time units.

The mutation operation swaps task t9 on processor p1 to task t8 on processor p3.

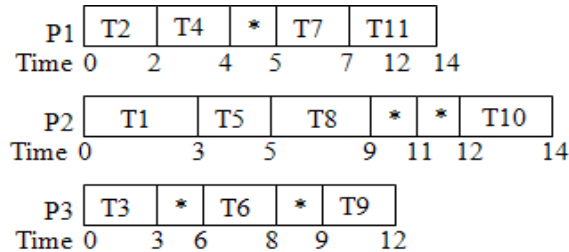


FIGURE 9(d): A Gantt chart after swap mutation operation

Makespan = 14 time units

Flowtime = 3+2+3+4+5+8+7+9+12+14+12 = 79 time units.

The mutation operation swaps task t8 on processor p1 to task t11 on processor p2 that takes 14 time units to complete.

The procedure of the Suggested Heuristics based Genetic Algorithm is:

Step 1: Setting the parameter

Set the parameter: Read DAG (number of tasks n , number of processors m and comm. cost), population size pop_size , crossover probability pc , mutation probability pm , and maximum generation $maxgen$.

Let generation $gen = 0$

Step 2: Initialization

Generate pop_size chromosomes using minimum execution time (MET)/Min-Min heuristic and b-level/t-level precedence resolutions.

Step 3: Evaluate

Calculate the fitness value of each chromosome

Step 4: Crossover

Perform the crossover operation on the chromosomes selected with probability pc .

Step 5: Mutation

Perform the swap/move mutation on chromosomes selected with probability pm .

Step 6: Selection

Select pop_size chromosomes from the parents and offspring for the next generation.

Step 7: Stop testing

If $gen = maxgen$, then output best solution and stop

Else $gen = gen + 1$ and return to step 3

4. EXPERIMENTAL RESULTS & PERFORMANCE ANALYSIS

The final best schedule obtained by applying the suggested algorithm to the DAG of figure 2 onto the parallel multiprocessor system in figure 1, is shown in figure 10 & 11. The completion time obtained by heuristics based method using b-level resolution is 14 time units and with t-level resolution is 16 time units. We also compare the results with FCFS scheduling method, min-min

scheduling method, MET scheduling method and also with the Basic Genetic Algorithm (BGA) [7] on parallel systems and execution of the schedule are shown in figure 12, 13, 14 & 15.

After applying the suggested heuristics based GA, the best schedule found using b-level precedence resolution is:

P1: T2→T4→T7→T11

P2: T1→T5→T8→T10

P3: T3→T6→T9

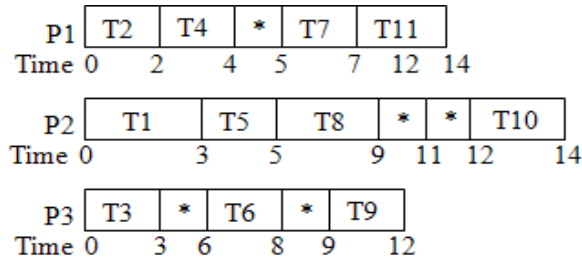


FIGURE 10: A Gantt chart of Suggested Heuristics based Genetic Algorithm using b-level resolution.

Makespan = 14 time units

Flowtime = 3+2+3+4+5+8+7+9+12+14+12 = 79 time units.

After applying the suggested heuristics based GA, the best schedule found using t-level precedence resolution is:

P1: T2→T4→T7→T11

P2: T1→T5→T10→T8

P3: T3→T6→T9

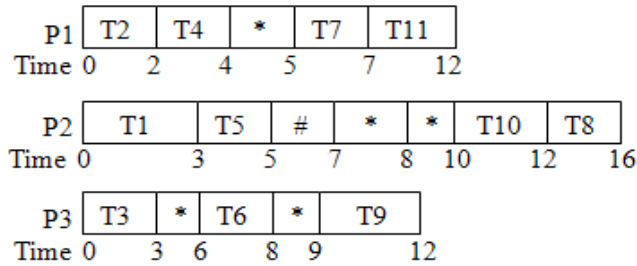


FIGURE 11: A Gantt chart of Suggested Heuristics based Genetic Algorithm using t-level resolution.

Makespan = 16 time units

Flowtime = 3+2+3+4+5+8+7+16+12+12+12 = 84 time units.

Min-min scheduling policy assigns the tasks to processors p1, p2 & p3 as:

P1: T2→T4→T6→T11

P2: T1→T5→T8→T10

P3: T3→T7→T9

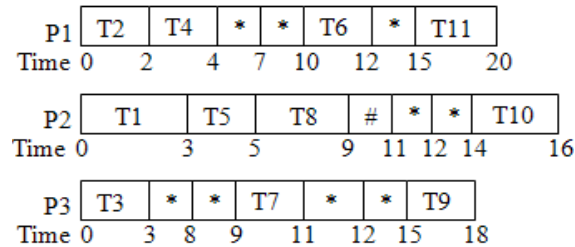


FIGURE 12: A Gantt chart of Min-min scheduler

Makespan = 20 time units

Flowtime = 3+2+3+4+5+12+11+9+18+16+20 = 103 time units.

FCFS scheduling Policy assign the tasks to processors p1, p2 & p3 as:

P1: T1→T4→T7→T10

P2: T2→T5→T8→T11

P3: T3→T6→T9

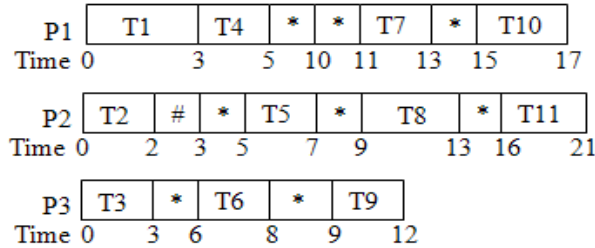


FIGURE 13: A Gantt chart of FCFS scheduler.

Makespan = 21 time units

Flowtime = 3+2+3+5+7+8+13+13+12+17+21 = 104 time units.

Minimum Execution Time (MET) Scheduling Policy assigns the tasks to processors p1, p2 & p3 as:

P1: T2→T5→T7→T8

P2: T4→T3→T10→T11

P3: T1→T6→T9

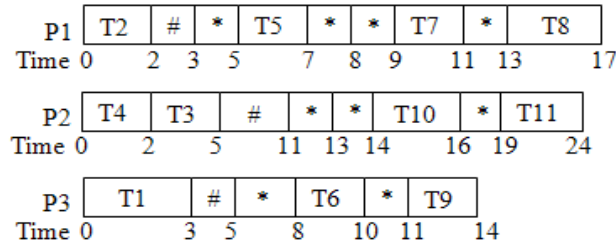


FIGURE 14: A Gantt chart of MET scheduler.

Makespan = 24 time units

Flowtime = 3+2+5+2+7+10+11+17+14+16+24 = 111 time units.

After applying the Basic GA, the best schedule found is:

P1: T1→T4→T7→T10

P2: T2→T5→T8→T11

P3: T3→T6→T9

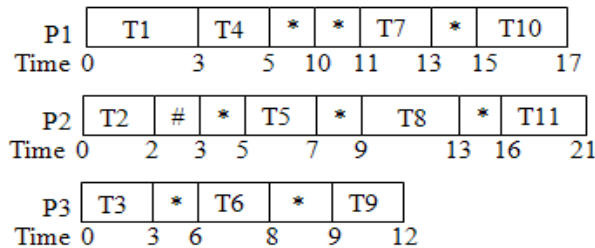
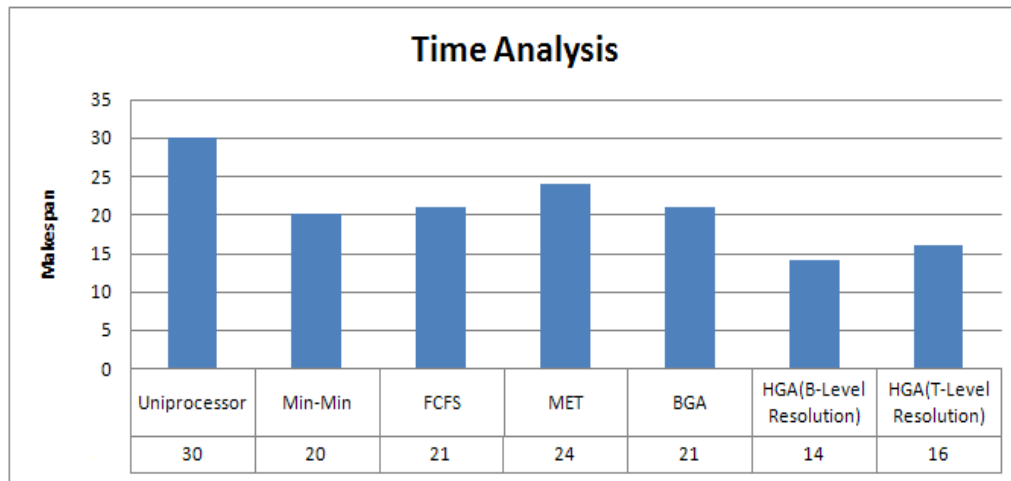


FIGURE 15: A Gantt chart of BGA scheduler.

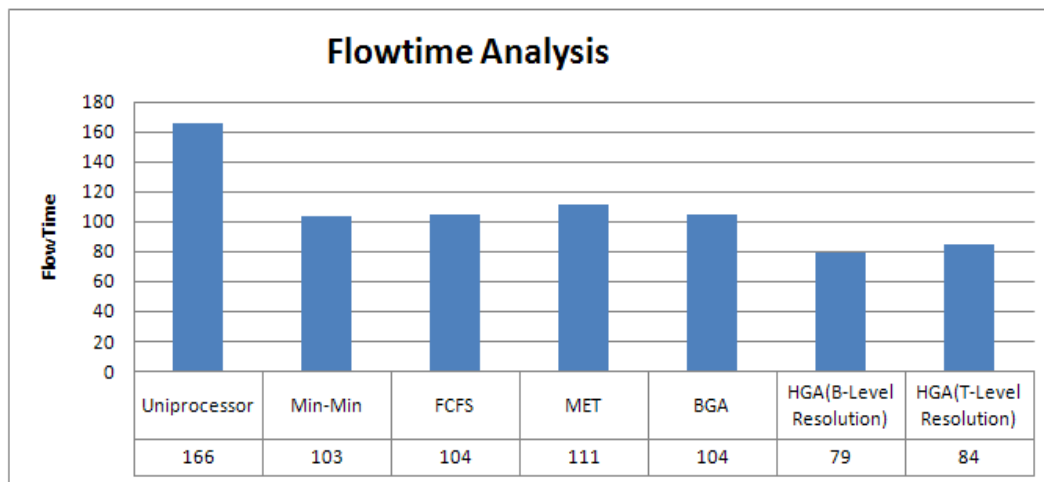
Makespan = 21 time units

Flowtime = 3+2+3+5+7+8+13+13+12+17+21 = 104 time units.

In Figure 16 (a) & (b), it is clear that HGA can considerably decrease the scheduling time.



(a)



(b)

FIGURE 16: Experimental results for (a) Makespan (b) Flowtime

Performance Analysis

1. Suggested Heuristics based GA using b-level resolution:

Speed up (S): speed up is defined as the completion time on a uniprocessor divided by completion time on a multiprocessor system.

$$S = 30/14 \\ = 2.142$$

Efficiency (E): $(S * 100) / m$, where m is the number of processors.

$$E = (2.142 * 100) / 3 = 71.42 \%$$

2. Suggested Heuristics based GA using t-level resolution

$$S = 30/16 = 1.875$$

$$E = (1.875 * 100) / 3 = 62.5 \%$$

3. Min-min Scheduler:

$$S = 30/20 = 1.5$$

$$E = (1.5 * 100) / 3 = 50 \%$$

4. FCFS Scheduler:

$$S = 30/21 = 1.428$$

$$E = (1.428 * 100) / 3 = 47.61 \%$$

5. MET Scheduler:

$$S = 30/24 = 1.25$$

$$E = (1.25 * 100) / 3 = 41.66 \%$$

6. BGA Scheduler:

$$S = 30/21 = 1.428$$

$$E = (1.428 * 100) / 3 = 47.61 \%$$

The performance analysis of various scheduling schemes is shown in figure 17.

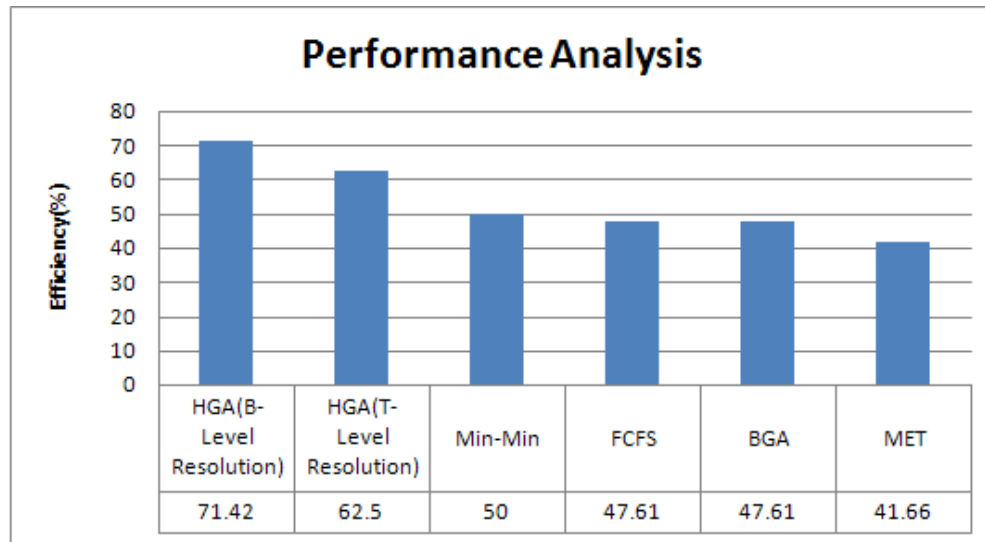


FIGURE 17: Performance analysis of min-min, FCFS, BGA, MET, HGA algorithms.

5. CONCLUSION

The task scheduling problem in the distributed systems is known to be NP-hard. The heuristic algorithms which obtain near-optimal solution in an acceptable interval time are preferred to the back tracking and the dynamic programming. The genetic algorithm is one of the heuristic algorithms which have the high capability to solve the complicated problems like the task scheduling.

In this paper, a new genetic algorithm, named Heuristics based Genetic Algorithm for Scheduling Static Tasks in Homogeneous parallel System is presented which its population size and the number of generations depends on the number of tasks. This algorithm tends to minimize the completion time and increase the throughput of the system. The heuristics based method found a best solution for assigning the tasks to the homogeneous parallel multiprocessor system. Experimental results and performance of the heuristics based GA with different precedence resolution methods is compared with Min-min, MET, FCFS and BGA Scheduling method and shows the efficiency of 71.42 %. The performance study is based on the best randomly generated schedule of the suggested GA.

6. REFERENCES

- [1] Ishfaq Ahmad, Yu-Kwong Kwok, Min-You Wu, "Analysis, Evaluation, and Comparison of Algorithms for Scheduling Task Graphs on Parallel Processors", Proceedings of the 1996 International Symposium on Parallel Architectures, Algorithms and Networks, Page: 207, 1996, ISBN: 0-8186-7460-1, IEEE Computer Society Washington, DC, USA.
- [2] Yu-Kwong Kwok and Ishfaq Ahmad, "Static Scheduling Algorithms for Allocating Directed Task Graphs to Multiprocessors", ACM Computing Surveys, vol. 31, Issue. 4, December 1999, ISSN: 0360-0300, ACM New York, NY, USA.
- [3] D. E. Goldberg, "Genetic algorithms in search, optimization & machine learning", Addison Wesley, 1990.

- [4] Melanie Mitchell, "An Introduction to Genetic algorithms", The MIT Press, February 1998.
- [5] Tracy D. Braunt, Howard Jay Siegel, Noah Beck, Bin Yao, Richard F. Freund, "A Comparison Study of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems", *Journal of Parallel and Distributed Computing*, Volume 61, Issue 6, June 2001, Pages: 810-837, ISSN: 0743-7315, Academic Press, Inc. Orlando, FL, USA.
- [6] Ricardo C. Correa, Afonso Ferreira, Pascal Rebreyend, "Scheduling Multiprocessor Tasks With Genetic Algorithms", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 10, Issue. 8, August 1999, Pages: 825-837, ISSN: 1045-9219, IEEE Press Piscataway, NJ, USA.
- [7] Edwin S. H. Hou, Nirwan Ansari, Hong Ren, "A Genetic Algorithm for Multiprocessor Scheduling", *IEEE Transactions on Parallel and Distributed Systems*, vol. 5, Issue. 2, February 2003, Pages: 113-120, ISSN: 1045-9219, IEEE Press Piscataway, NJ, USA.
- [8] Amir Masoud Rahmani, Mohammad Ali Vahedi, "A novel Task Scheduling in Multiprocessor Systems with Genetic Algorithm by using Elitism stepping method", *Science and Research branch*, Tehran, Iran, May 26, 2008.
- [9] Martin Grajcar, "Genetic List Scheduling Algorithm for Scheduling and Allocating on a Loosely Coupled Heterogeneous Multiprocessor System", *Proceedings of the 36th annual ACM/IEEE Design Automation Conference*, New Orleans, Louisiana, United States, Pages: 280 – 285, 1999, ISBN: 1-58133-109-7, ACM New York, NY, USA.
- [10] Martin Grajcar, "Strengths and Weaknesses of Genetic List Scheduling for Heterogeneous Systems", *IEEE Computer Society, Proceedings of the Second International Conference on Application of Concurrency to System Design*, Page: 123, ISBN: 0-7695-1071-X, IEEE Computer Society Washington, DC, USA, 2001.
- [11] Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Comparison of Heuristics for scheduling Independent Tasks on Heterogeneous Distributed Environments", *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization*, Volume 01, Pages: 8-12, 2009, ISBN:978-0-7695-3605-7, IEEE Computer Society Washington, DC, USA.
- [12] Yi-Hsuan Lee and Cheng Chen, "A Modified Genetic Algorithm for Task Scheduling in Multiprocessor Systems", *Proc. of 6th International Conference Systems and Applications*, pp. 382-387, 1999.
- [13] Amir Masoud Rahmani and Mojtaba Rezvani, "A Novel Genetic Algorithm for Static Task Scheduling in Distributed Systems", *International Journal of Computer Theory and Engineering*, Vol. 1, No. 1, April 2009, 1793-8201.
- [14] Michael Rinehart, Vida Kianzad and Shuvra S. Bhattacharyya, "A modular Genetic Algorithm for Scheduling Task Graphs", *Technical Report UMIACS-TR-2003-66*, Institute for Advanced Computer Studies University of Maryland at College Park, June 2003.
- [15] Pai-Chou Wang, W. Korfhage, "Process Scheduling with Genetic Algorithms", *Proceedings of the 7th IEEE Symposium on Parallel and Distributed Processing*, Page: 638, ISBN: 0-8186-7195-5, October 2005, IEEE Computer Society Washington, DC, USA.
- [16] Prof. Sanjay R Sutar, Jyoti P. Sawant, Jyoti R. Jadhav, "Task Scheduling For Multiprocessor Systems Using Memetic Algorithms", <http://www.comp.brad.ac.uk/het-net/tutorials/P27.pdf>
- [17] Andrew J. Page and Thomas J. Naughton, "Framework for Task scheduling in heterogeneous distributed computing using genetic algorithms", *15th Artificial Intelligence and Cognitive Science Conference*, 2004, Castlebar, Co. Mayo, Ireland, isbn = 1-902277-89-9 pages = 137-146.
- [18] Clayton S. Ferner and Robert G. Babb, "Automatic Choice of Scheduling Heuristics for Parallel/Distributed Computing", *IOS Press Amsterdam, The Netherlands*, Volume 7, Issue 1, Pages: 47 – 65, January 1999, ISSN:1058-9244.
- [19] C.L. McCreary, A.A. Khan, J. Thompson, M.E. McArdle, "A Comparison of Heuristics for Scheduling DAGs on Multiprocessors", *Eighth International Proceedings on Parallel Processing Symposium*, pages: 446-451, Location: Cancun, ISBN: 0-8186-5602-6, DOI: 10.1109/IPPS.2002.288264, 06 August 2002.
- [20] Javier Carretero, Fatos Xhafa, Ajith Abraham, "Genetic Algorithm Based Schedulers for Grid Computing Systems", *International Journal of Innovative Computing, Information and Control*, ICIC International, Vol.3, No. 6, ISSN 1349-4198, pp. 1053-1071, December 2007.

- [21] Annie s. Wu, Han Yu, Shiyuan Jin, Kuo-Chi Lin, and Guy Schiavone, "An Incremental Genetic Algorithm Approach to Multiprocessor Scheduling", IEEE Transactions on Parallel and Distributed Systems, Vol.15, No. 9, On page(s): 824 – 834, ISSN: 1045-9219, INSPEC Accession Number:8094176, Digital Object Identifier: 10.1109/TPDS.2004.38, 13 September 2004.
- [22] Michael Bohler, Frank Moore, Yi Pan, "Improved Multiprocessor Task Scheduling Using Genetic Algorithms", Proceedings of the Twelfth International FLAIRS Conference, WPAFB, OH 45433, American Association for Artificial Intelligence, 1999.
- [23] Marin Golub, Suad Kasapovic, "Scheduling Multiprocessor Tasks with Genetic Algorithms", OACTA Press, from proceeding (351) Applied Informatics, 2002.
- [24] M. Nikravan and M. H. Kashani, "A Genetic Algorithm for Process Scheduling in Distributed Operating Systems considering Load Balancing", Proceedings 21st European Conference on Modelling and Simulation Ivan Zelinka, Zuzana Oplatkova, Alessandra Orsoni, ECMS 2007, ISBN 978-0-9553018-2-7, ISBN 978-0-9553018-3-4 (CD).
- [25] Shuang E Zhou, Yong Liu, Di Jiang, "A Genetic-Annealing Algorithm for Task Scheduling Based on Precedence Task Duplication", CIT, Proceedings of the Sixth IEEE International Conference on Computer and Information Technology, Page: 117, 2006, ISBN: 0-7695-2687-X, IEEE Computer Society Washington, DC, USA.

Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach

Oladipupo O.O.

*College of Science and Technology, Department
of Computer and Information Sciences
Covenant University
Ota, PMB 1023, Nigeria*

frajooje@yahoo.com

Oyelade O.J.

*College of Science and Technology, Department
of Computer and Information Sciences
Covenant University
Ota, PMB 1023, Nigeria*

ayo2006ola@yahoo.com

Abstract

Over the years, several statistical tools have been used to analyze students' performance from different points of view. This paper presents data mining in education environment that identifies students' failure patterns using association rule mining technique. The identified patterns are analysed to offer a helpful and constructive recommendations to the academic planners in higher institutions of learning to enhance their decision making process. This will also aid in the curriculum structure and modification in order to improve students' academic performance and trim down failure rate. The software for mining student failed courses was developed and the analytical process was described.

Keyword: Association rule mining, Academic performance, Educational data mining, Curriculum, Students' Result Repository.

1. INTRODUCTION

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process [1]. It is concerned with developing methods for exploring the unique types of data that come from educational environment which include students' results repository.

Students' result repository is a large data bank which shows the students raw scores and grades in different courses they enrolled for during their years of attendance in the institution. Student performance score is basically determined by the sum total of the continuous assessment and the examination scores. In most institutions the continuous assessment which includes various assignments, class tests, group presentations is summed up to weigh 30% of the total score while the main semester examination is 70%. To differentiate different students' performances and scores a set of alphabetic grade is identified to represent the score ranges such as 70-100

as “A”, 60-69 as “B”, 50 to 59 as “C” and 45-49 as “ D” and < 45 as “F”. Any score < 45 is regarded as a fail performance. This grade representation is different from one higher institution to another.

From the standpoint of the e-learning scholars, data mining techniques is said to have been applied to solve different problems in educational environment which includes Students' classification based on their learning performance; detection of irregular learning behaviors; e-learning system navigation and interaction optimization; clustering according to similar e-learning system usage; and systems' adaptability to students' requirements and capacities and so on. [2] The choice of data mining tool is mostly determined by the scope of the problem and the expected analysis result.

In [3] an approach to classify students in order to predict their final year grade based on the features extracted from logged data in an educational web-based system was reported. Data mining classification process was used in conjunction with genetic algorithm to improve the prediction accuracy. Also, in [4] student data was mined to characterize similar behavior groups in unstructured collaboration using clustering algorithms. The relationship between students' university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques in [5]. Fuzzy logic concept was not behind in the field of educational data mining [6,7,8,9], for instance a two-phase fuzzy mining and leaning algorithm was described in [10], this is an hybrid system of association rule mining apriori algorithm with fuzzy set theory and inductive learning algorithm to find embedded information that could be fed back to teachers for refining or reorganizing the teaching materials and test. Association rule mining technique has also been used in several occasions in solving educational problems and to perform crucial analysis in the educational environment. This is to enhance educational standards and management such as investigating the areas of learning recommendation systems, learning material organization, student assessments, course adaptation to the students' behaviour and evaluation of educational web sites [1,11,12 13,14]. In [12] a Test Result Feedback (TRF) model that analyses the relationships between students' learning time and the corresponding test results was introduced. Knowledge Discovery through Data Visualization of Drive Test Data was carried out in [15]. Genetic algorithm as Ai technique was for data quality mining in [16] Association rule mining was used to mining spartial Gene Expressing [17] and to discover patterns from student online course usage in [14] and it is reported that the discovered patterns from student online course usage can be used for the refinement of online course. Robertas, in [18] analysed student academic results for informatics course improvement, rank course topics following their importance for final course marks based on the strength of the association rules and proposed which specific course topic should be improved to achieve higher student learning effectiveness and progress.

In view of the literature, it is observed that different analysis has been done on students' result repository but the failed courses in isolation has never been analysed for hidden and important patterns, which could be of a great importance to academic planners in enhancing their decision making process and improving student performance. In order to bridge this gap, this paper presents an analysis of students' academic failed courses in isolation using association rule mining. This is to discover the hidden relationships that exist between different students failed courses in form of rules. The generated rules are analysed to make useful and constructive recommendations to the academic planners. This promised to enhance academic planner's sense of decision making and aid in the curriculum structure and modification which in turn improve students' performance and trim down failure rate.

1.1 Association Rule mining

Association rule mining associates one or more attributes of a dataset with another attributes, to discover hidden and important relationship between the attributes, producing an if-then statement concerning attribute values in form of rules. (19,20). The formal definition of association rule

mining is : Let $I = \{i_1, i_2, \dots, i_m\}$ be a set of literals called items and D be a set of transactions where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X , a set of some items in I , if $X \subseteq T$. Association rule mining process could be decomposed into two main phases to enhance the implementation of the algorithm. The phases are:

1. Frequent Item Generation: This is to find all the itemsets that satisfy the minimum support threshold. The itemsets are called frequent itemsets.
2. Rule Generation: This is to extract all the high confidence rules from the frequent itemsets found in the first step. These rules are called strong rules.

Over the years different algorithms have been proposed in the literature that implement the two phases of association rule mining [21]. In this paper the traditional Apriori algorithm is implemented to generate the hidden patterns from the students' failed courses dataset which when analysed will serve as a strong convincing recommendation to academic planning department in institutions of learning for curriculum structure and modification in other to improve the students' performances and minimize failure rate percentage.

1.2 Definition of terms

Association rules: An association rule is an implication expression of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and X and Y are disjoint itemsets, i.e $X \cap Y = \phi$. The strength of an association rule can be measured in terms of its support and confidence. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c and support s , if $c\%$ of the transactions in D that contains X also contains Y , and $s\%$ of transactions in D contains $X \cup Y$. Both the antecedent and the consequent of the rule could have more than one item. The formal definitions of these two metrics are:

$$\text{Support, } s(X \Rightarrow Y) = \frac{\sum(X \cup Y)}{N} \quad (1)$$

$$\text{Confidence, } c(X \Rightarrow Y) = \frac{\sum(X \cup Y)}{\sum X} \quad (2)$$

Example 1: Consider a rule $\{CSC111, CSC121\} \Rightarrow \{CSC211\}$. If the support count for $\{CSC111, CSC121, CSC211\}$ is 2 and the total number of transactions is 5 then, the rule's support is $2/5 = 0.4$. The rule's confidence is obtained by dividing the support count for $\{CSC111, CSC121, CSC211\}$ by the support count for $\{CSC111, CSC121\}$. If there are 3 transactions that contain $CSC111, CSC121$ then, the confidence for this rule is $2/3 = 0.67$. If the minimum rule support is 0.3 and minimum confidence is 0.5, then, the rule $\{CSC111, CSC121\} \Rightarrow \{CSC211\}$ is said to be strong, that is; the interestingness of the rule is high.

1.3 Justification for Support and Confidence measure

Support is an important measure because a rule that has a very low support may occur by chance. A low support rule in this context is likely to be uninteresting from the academic perspective because such a failure combination might come accidentally and it might not be profitable to enhance academic planner decision. Also, confidence on the other hand, measures the reliability of the inference made by a rule. So, the higher the confidence, the more frequent the failed courses appear together within the database.

2. METHODOLOGY:

2.1 Development of an Apriori Algorithm

The algorithm starts by collecting all the frequent 1-itemsets in the first pass based on the minimum support. It uses this set (called L_1) to generate the candidate sets to be frequent in the next pass (called C_2) by joining L_1 with itself. Any item that is in C_1 and not in L_1 is eliminated from C_2 . This is achieved by calling a function called 'apriori-gen'. This reduces the item size drastically. The algorithm continues in the same way to generate the C_k , of size k from the large itemsets of $k-1$, then reduces the candidate set by eliminating all those items in $k-1$ with support count less than minimum support. The algorithm terminates when there are no candidates to be counted in the next pass. Figure 1 shows the general Pseudocode for association rule mining and Figure 2 shows the traditional apriori algorithm, while figure 3 , shows the algorithm for 'apriori-gen' function called for candidate generation and elimination of non-frequent itemset.

| | |
|----------|---|
| Step 1: | Accept the minimum support as minsup and minimum confidence as minconf and the student failed course as the input data set. |
| Step 2: | Determine the support count for all the item as s (courses under consideration). |
| Step 3: | Select the frequent items; item with $s \geq \text{minsup}$ |
| Step 4: | The set candidate k - item is generated by 1- extension of the large $(k-1)$ itemsets generated in step3 |
| Step 5: | Support for the candidate k -itemsets are generated by a pass over the database. |
| Step 6; | Itemset that do not have minsup are discarded and the remaining itemsets are called large k -itemsets. |
| Step 7 : | The process is repeated until no more large item. |
| Step 8: | The interesting rules are determined based on the minimum confidence. |

FIGURE 1: General Pseudocode for Association Rule Mining

2.2 Apriori Candidate Generation

The apriori-gen is a function called in line 3 of the algorithm1. It takes as argument L_{k-1} , the set of all large $(k-1)$ itemsets. It returns a superset of the set of all large k -itemsets. The description of the function is given in algorithm 2.

2.3 Rule generation

After all the frequent itemsets have been generated then the rules are determined. In rule generation, we do not have to make additional passes over the data set to compute the support of the candidate rules. All needed is to determine the confidence of each rule by using the support counts computed during frequent itemsets generation.

Find frequent set L_{k-1} .

Join Step.

C_k is generated by joining L_{k-1} with itself

Prune Step.

Any $(k-1)$ -itemset that is not frequent cannot be a subset of a frequent k -itemset, hence should be removed.

where

(C_k : Candidate itemset of size k)

(L_k : frequent itemset of size k)

FIGURE 2: Frequent itemset generation of the Traditional Apriori Algorithm [21]

Apriori (T, ϵ)

$L_1 \leftarrow \{ \text{large 1-itemsets that appear in more than } \epsilon \text{ transactions} \}$

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \text{Generate}(L_{k-1})$

for transactions $t \in T$

$C_t \leftarrow \text{Subset}(C_k, t)$

for candidates $c \in C_t$

count[c] \leftarrow count[c] + 1

$L_k \leftarrow \{ c \in C_k \mid \text{count}[c] \geq \epsilon \}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

FIGURE 3: Function for generating candidate itemset [21]

3. Result and Rule analysis

In most literature authors focus on the students' aggregate performances; Grade Point Average [12,13,14, 18] and their findings are useful majority for prediction, which might not really improve the low capacity students' performances. In this research the association rule mining analysis was performed based on students' failed courses. This identifies hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. Figure 4 shows a snapshot for association mining process interface. In this work it is observed that the lower the items minimum support, the larger the candidate generated. This adversely affects the complexity of the system. For instance, in figure 4, if the item minimum

support is 3 and the rule confidence is 0.5, we have 19 frequent itemsets and 114 rules are generated. Table 2 show the relationship between the minimum support, minimum confidence and the generated rule and figure 5 gives the graphical representation.

It was observed that the execution time is also inversely proportional to minimum support, since it increases as minimum support decreases, which confirmed increase in system complexity and response time as the minimum support decreases as shown on table 2. With all these observations it shows that to have a less complex system and a constructive, interesting and relevant patterns the minimum confidence and support should be large enough to trash out coincidence patterns. Table 3 also displayed some of the rules generated.

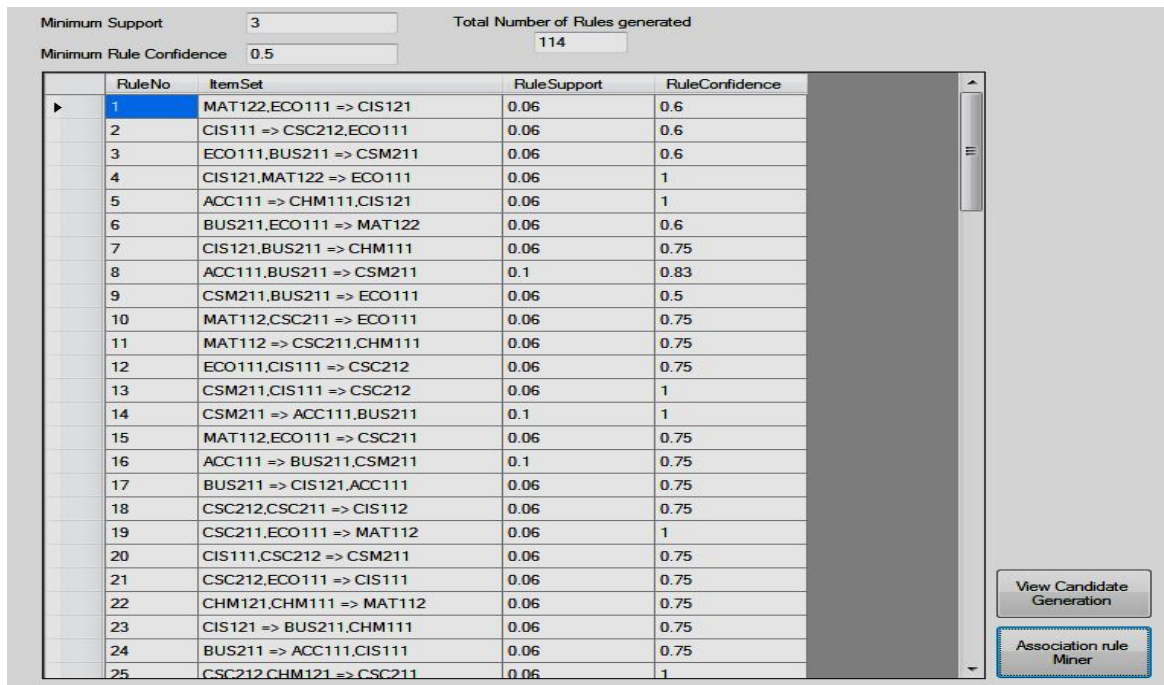


FIGURE 4 :A snapshot for Association rule mining process interface

TABLE 1: Relationship between minimum confidence , minimum support and number of generated rules.

| Min.Conf. | <i>Minimum Item(s) support = 3 Average Exe.Time = 6sec</i> | | <i>Minimum item(s) support = 3 Average Exe.Time = 14sec</i> | | <i>Minimum item(s) support = 3 Average Exe.Time = 40sec</i> | |
|-----------|--|-------------------|---|-------------------|---|-------------------|
| | #Rules | #of freq. Itemset | #Rules | #of freq. Itemset | #Rules | #of freq. Itemset |
| 50% | 6 | 1 | 114 | 19 | 855 | 152 |
| 60% | 6 | 1 | 97 | 19 | 631 | 152 |
| 70% | 6 | 1 | 82 | 19 | 345 | 152 |
| 80% | 6 | 1 | 51 | 19 | 306 | 152 |
| 90% | 2 | 1 | 49 | 19 | 301 | 152 |
| 100% | 2 | 1 | 49 | 19 | 301 | 152 |

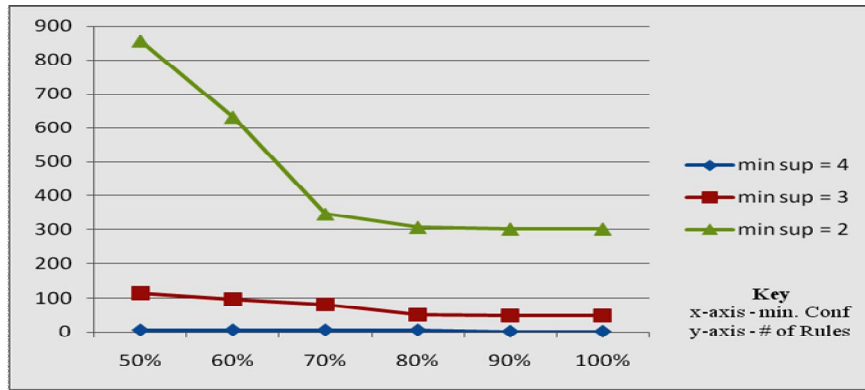


FIGURE 3: Graphical representation of effect of minsup, minconf on #of rules

3.1 Rule Analysis

Table 2 shows an instance of the rule generated from the simulation. All the rules with confidence 1, are very strong rules, which implies that if a student failed the determinant (antecedent) course(s), such student will surely fail the dependent (consequent) course(s). Such rules should not be overlooked in curriculum structure. Also if the rule support is higher, it means that all the courses involved are failed together by most of the considered students. From rule number 1 one can deduce that MAT 122, ECO 111 ⇒ CIS 121 with (s = 0.06, c = 0.6). This indicates that, the probability that every student that fails MAT 122, ECO 111 will also fail CIS 121 is 0.6. This type of rule is not very strong; in some cases it might be overlooked but notwithstanding, the academic planner can still take it into consideration. In that case, MAT 122 and CIS 121 should not be taken in the same semester. This kind of failure can be minimized if one becomes a prerequisite to another. That is, if a student has not passed MAT 122 they will not be allowed to register for CIS 121. Also, we have from rule 8, a strong rule such that ACC111, BUS211 ⇒ CSM211 with (s = 0.1, c = 0.83). ACC 111 is introduction to accounting; BUS 211 is introduction to Business and CSM 211 Mathematical method 1. The first two courses are compulsory courses for the Management Information System students. ACC111 is a 100 level first semester course while the other two are 200 level first semester courses. This implies that a student that fails ACC 111 in 100 level should not be allowed to register for BUS 211 or CSM 211 and if possible, the two, so as to avoid multiple failure.

With all these observations, if academic planners can make use of the extracted hidden patterns from students' failed causes using association rule mining approach, it will surely help in curriculum re-structuring and also, help in monitoring the students' ability. This will enable the academic advisers to guide students properly on courses they should enroll for. This, eventually, tends to increase the student pass rate.

TABLE 2: An instance of rule generated with support and confidence

| RuleNO | Rule | Rule Support | Confidence |
|--------|------------------------|--------------|------------|
| 1 | MAT122,ECO111 ⇒ CIS121 | 0.06 | 0.6 |
| 2 | CIS111 ⇒ CSC212,ECO111 | 0.06 | 0.6 |
| 4 | CIS121,MAT122 ⇒ ECO111 | 0.06 | 1 |
| 5 | ACC111 ⇒ CHM111,CIS121 | 0.06 | 1 |
| 6 | BUS211,ECO111 ⇒ MAT122 | 0.06 | 0.6 |
| 7 | CIS121,BUS211 ⇒ CHM111 | 0.06 | 0.75 |
| 8 | ACC111,BUS211 ⇒ CSM211 | 0.1 | 0.83 |

4. Conclusion, Recommendation and Future Work

This study has bridge the gap in educational data analysis and shows the potential of the association rule mining algorithm for enhancing the effectiveness of academic planners and level advisers in higher institutions of leaning. The analysis was done using undergraduate students' result in the department of Computer Science from a university in Nigeria. The department offers two programmes; Computer Science and Management Information Science. A total number of 30 courses for 100 level and 200 level students are considered as a case study. The analysis reveals that there is more to students' failure than the students' ability. It also reveals some hidden patterns of students' failed courses which could serve as bedrock for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate. To adopt this approach a larger number of students should be considered from the first year to the final year in the institution. This will surely reveal more interesting patterns. Also, the min. confidence should be of a higher percentage to be able to have more relevant and constructive rules. In future applications, in order to improve the comprehensibility and applicability of the association rules, it will be very useful to also provide an ontology that would describe the content of the courses which will allow the academic planners to understand better the rules that contain concepts related to the analysed domain.

References

1. B. Dogan, A. Y. Camurcu. "Association Rule Mining from an Intelligent Tutor" Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 447, 2008
2. F. Castro, A. Vellido, A. Nebot, and F. Mugica. "Applying Data Mining Techniques to e-Learning Problems". Evolution of Teaching and Learning Paradigms in Intelligent Environment ISBN: 10.1007/978-3-540-71974-8_8 Volume 62, pp 183-221. Springer Berlin Heidelberg, 2007.
3. B.Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and, W. F. Punch."Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA" In Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.
4. Talavera, L., and Gaudio, E. "Mining student data to characterize similar behavior groups in unstructured collaboration spaces". In Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI ,Valencia, Spain, 2004.
5. Ş. Z. ERDOĞAN, M. TİMOR . "A data mining application in a student database". Journal of aeronautics and space technologies ,volume 2 number 2 (53-57) 2005.
6. G.J. Hwang. "A Knowledge-Based System as an Intelligent Learning Advisor on Computer Networks" Journal of Systems, Man, and Cybernetics Vol. 2 , pp.153-158, 1999.
7. G.J. Hwang, T.C.K. Huang,and C.R. Tseng. "A Group-Decision Approach for Evaluating Educational Web Sites". Computers & Education Vol. 42 pp. 65-86 , 2004.
8. G.J. Hwang, C.R. Judy, C.H. Wu, C.M. Li and G.H. Hwang. "Development of an Intelligent Management System for Monitoring Educational Web Servers". In proceedings of the 10th Pacific Asia Conference on Information Systems, PACIS . 2334-2340, 2004.
9. G.D. Stathacopoulou, M. Grigoriadou. "Neural Network-Based Fuzzy Modeling of the Student in Intelligent Tutoring Systems". In proceedings of the International Joint Conference on Neural Networks. Washington ,3517-3521,1999.

10. C.J. Tsai, S.S. Tseng, and C.Y. Lin. "A Two-Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment". In proceedings of the Alexandrov, V.N., et al. (eds.): International Conference on Computational Science, ICCS 2001. LNCS Vol. 2074. Springer-Verlag, Berlin Heidelberg New York, 429-438. 2001.
11. S. Encheva, S. Tumin. "Application of Association Rules for Efficient Learning Work-Flow" Intelligent Information Processing III, ISBN 978-0-387-44639-4, pp 499-504 published Springer Boston, 2007.
12. H.H. Hsu, C.H. Chen, W.P. Tai. "Towards Error-Free and Personalized Web-Based Courses". In proceedings of the 17th International Conference on Advanced Information Networking and Applications, AINA'03. March 27-29, Xian, China, 99-104, 2003.
13. P. L. Hsu, R. Lai, C. C. Chiu, C. I. Hsu (2003) "The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance" [Expert Systems with Applications 25 (2003) 51–62.
14. A.Y.K. Chan, K.O. Chow, and K.S. Cheung. "Online Course Refinement through Association Rule Mining" Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 44, 2008.
15. S. Saxena, A. S.Pandya, R. Stone, S. R. and S. Hsu (2009) "Knowledge Discovery through Data Visualization of Drive Test Data" International Journal of Computer Science and Security (IJCSS), Volume (3): Issue (6) pp. 559-568.
16. S. Das and B. Saha (2009) "Data Quality Mining using Genetic Algorithm" International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (2) pp. 105-112
17. M.Anandhavalli, M.K.Ghose and K.Gauthaman(2009) "Mining Spatial Gene Expression Data Using Association Rules". International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (5) pp. 351-357
18. R. Damaševicius. "Analysis of Academic Results for Informatics Course Improvement Using Association Rule Mining". Information Systems Development Towards a Service Provision Society. ISBN 978-0-387-84809-9 (Print) 978-0-387-84810-5 (Online) pp 357-363, published by Springer US, 2009.
19. Ceglar, J.F Roddick. "Association mining". ACM Computing Surveys, 38:2, pp. 1-42, 2006
20. S. Kotsiantis, D. Kanellopoulos. "Association Rules Mining" A Recent Overview.GESTS Int. Transactions on Computer Science and Engineering, Vol. 32 (1), pp. 71-82, 2006.
21. H. Jochen, G. Ulrich and N. Gholamreza. "Algorithms for Association Rule Mining – A General Survey and Comparison". SIGKDD Exploration, Vol.2, Issue 1, pp 58-64. ACM, 2000.

Multilevel Access Control in a MANET for a Defense Messaging system using Elliptic Curve Cryptography

J. Nafeesa Begum

nafeesa_jeddy@yahoo.com

Sr.Lecturer/CSE

*Government College of Engineering, Bargur
Krishnagiri District, Tamil Nadu, India*

K.Kumar

pkk_kumar@yahoo.com

Lecturer/CSE

*Government College of Engineering, Bargur
Krishnagiri District, Tamil Nadu, India*

Dr.V.Sumathy

sumathy_gct2001@yahoo.co.in

AssistantProfessor/ECE

*Government College of Technology, Coimbatore
Coimbatore District, Tamil Nadu, India*

Abstract

The trend of the Civilian society has moved from the industrial age focus on automation and scale towards information based on computing and communication. Today's Warfare is also moving towards an information age paradigm based on information sharing, situational awareness, and distributed points of intelligence, command and control. A widely-networked fighting force is better able to share information about tactical situations that may be geographically widespread, asymmetric, and rapidly changing. Commanders must be able to better assess situations across broad theaters, with extensive data, voice, and especially video feeds as strategic inputs. Thus, network-centric warfare improves effectiveness at both the tactical "point of the spear" and in the achievement of broader strategic goals. Broadly disseminated knowledge assets enable fighting forces that must self-synchronize, even as they physically disperse to address dynamic battlefield conditions. The speed of decision has increased and command decisions must be rapidly relayed and implemented, to improve battlefield outcomes. Multilevel access control in a MANET for a Defense messaging system is used to have the command decisions relayed to all people who are active in the group and also to all people who have been identified as higher in the hierarchy instead of sending one to one messages to each individual.. The system developed is secure, multi site and allows for global communication using the inherent properties of Elliptic Curve cryptography . Elliptic Curve cryptography provides a greater security with less bit size and it is fast when compared to other schemes. The implementation suggests that it is a secure system which occupies fewer bits and can be used for low power devices.

Keywords: Defense messaging system, Elliptic Curve cryptography, Encryption , Global Information Sharing , Secure system.

1. INTRODUCTION

Information superiority has become as important in today's battlefield as air superiority was in the past in increasing mission effectiveness. Information superiority has become critical as needs of both war fighters and commanders have broadened to include real-time video, high-speed data, and voice. Data and intelligence sources include terrestrial forces and sensors, satellites, UAVs (Unmanned Aerial Vehicles), and a wide variety of centralized and distributed information assets.[7,8] The vast majority of these information assets, command, communications, and control must be delivered wirelessly, with seamless connections to wired networks for intelligence resources and other data. Further, these wireless technologies must support data, voice, and increasingly, video traffic flows. In the network-centric warfare environment, mobility implies more than just the motion of individuals and vehicles in relation to one another and to other fixed locations.

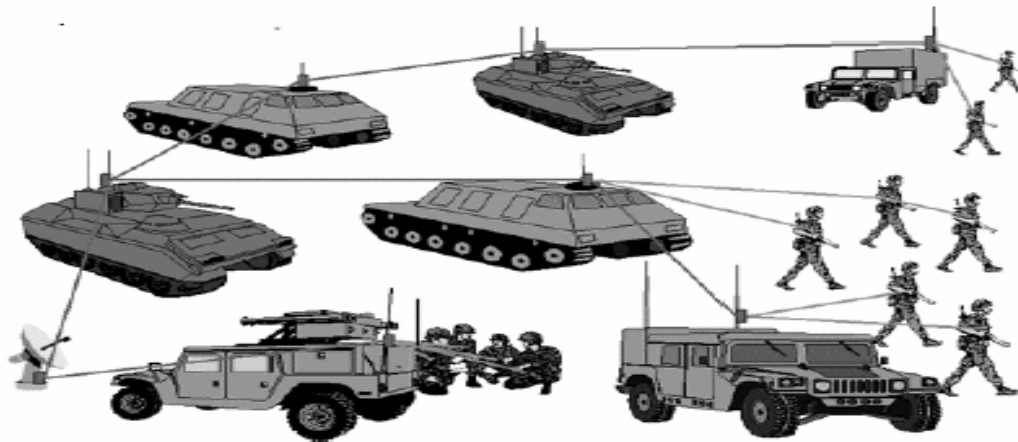


Figure:1 MANET In a military warfare

In such a mobile adhoc network a Defense messaging system takes a message and forwards it to the intending recipients or parties based on the message criteria for immediate action. It enables the top defense personnel to issue commands using messages to guard the country against threats of terrorists, anti socials and Intruders. The system developed has the following capabilities

- Encryption of the data stream
- Strong user authentication
- Prevention of interaction with external and undesirable applications
- Platform independent
- Central administration and logging

As a multilevel access control systems, the system developed has the following features

- Hiding of hierarchy and receivers
- Authentication of receivers
- Dynamics at message level, class level and user level by the server

The main advantage of ECC compared to other schemes is that it offers equal security with a smaller key size and thus reduces processing overhead and can be used in Tiny devices also. The rest of the paper is as follows. Section 2 gives the related work Section 3 describes Elliptic Curve Cryptography. Section 4 explains our system development, algorithms and dynamics. Section 5 gives the implementation results, Section 6 does the Performance Analysis and Section 7 concludes the paper.

2.RELATED WORK

The first multi level access solution was proposed by Akl et al.[1] in 1983 and followed by many others [8,9,10,11,13,14,15,16 ,12]. These schemes basically rely on a one-way function so that a node v can easily compute v 's descendants' keys whereas v 's key is computationally difficult to compute by v 's descendant nodes. Moreover, many existing schemes have some of the following problems: (1) Some schemes were found with security flaws shortly after they were proposed [3,4] (2) Some schemes cannot support for reconfiguration of a hierarchy [18,19]; (3) Some schemes require access hierarchy to be in a certain form so that it must be a tree or a DAG with only one root; and (4) Member revocation is one of the most difficult processes in cryptographic schemes, therefore, it is important to address this problem so that the revocation process can be dealt with efficiently. In this paper, we propose a new scheme based on elliptic curve cryptography for secret messaging which has the suitable characteristics. Unlike many existing schemes based on one-way functions, our scheme is based on a secret sharing method which makes the scheme unconditionally secure [17, 20]. Elliptic curve cryptography address the issue of saving the power due to the use of less number of bits for secure transmission [2]. In our previous work we have used [5,6] Elliptic curves used for efficient group key management. In this paper we have extended it to include multilevel access control. Multilevel access control using elliptic curve cryptography is a new research area under deployment and we have used it in the defense messaging system so that higher group members can see the messages relayed to lower group members.

3. ELLIPTIC CURVE CRYPTOGRAPHY

3.1 Basics

Elliptic curves are named so as they appear to be similar to the equation defining the roots of an ellipse. They are equations containing two variables and coefficients where the elements are in a finite field (\mathbb{Z}_p) . The elliptic equation is of the form $y^2 = x^3 + ax + b$. The coefficients a, b should satisfy the condition $4a^3 - 27b^2 \neq 0$ so that there are no repeated factors. For given values of a and b , the elliptic curve consists of positive and negative values of y for each value of x . A special point O which acts as an identity is used. The following addition rules are used in elliptic curve arithmetic.

1. $P + O = O + P = P$ for all P belongs to \mathbb{Z}_p
2. If $P = (x, y) \in E(\mathbb{Z}_p)$ then $(x, y) + (x, -y) = o$ and is called the negative of P and $-P$ is a point on the curve.
3. Let $P = (x_1, y_1) \in E(\mathbb{Z}_p)$ $Q = (x_2, y_2) \in E(\mathbb{Z}_p)$ where $P \neq Q$ then $P + Q$ is (x_3, y_3)
 $x_3 = (\lambda^2 - x_1 - x_2) \bmod P$
 $y_3 = (\lambda(x_1 - x_3) - y_1) \bmod P$ and
 $\lambda = ((y_2 - y_1) / (x_2 - x_1)) \bmod P$ if $P \neq Q$
 $\lambda = ((3x_1^2 + a) / 2y_1) \bmod P$ if $P = Q$
4. Multiplication is defined by repeated addition

In case of a finite group $E_p(a, b)$ the number of points on the elliptic curve is bounded by $P+1 - 2\sqrt{P} \leq N \leq P + \sqrt{P}$ so that for a large P , the number of elements is approximately equal to P .

3.2 Algorithm For Elliptic Curve Cryptography

Step1: Decide on the elliptic curve E . The Elliptic curve should have two coefficients a, b such that $4a^3 - 27b^2 \neq 0$ and p a prime number.

Step2: For the elliptic curve equation apply values of x from 1 to $p-1$ and generate y values.

Step3: Find the Quadratic residues to avoid repetition in mod values.

Step4: Collect all the points on the elliptic curve.

Step5 : Use one point called the base point as the generator using which scalar multiplications are performed and generate multiples of the generator by applying the ECC arithmetic rules.

Step 6: For a same elliptic curve, by choosing a different generator point we can obtain a different encryption values.

3.3. Example

Points in Elliptic Curve :The Points in $E_{211}(0,-4)$ are found using steps 1 to 5.

| | | | | | | | | | |
|--------|--------|-------|--------|--------|--------|-------|--------|--------|--------|
| 1,29 | 1,182 | 2,2 | 2,209 | 5,11 | 5,200 | 6,1 | 6,210 | 12,6 | 12,205 |
| 13,100 | 13,111 | 14,29 | 14,182 | 16,100 | 16,111 | 17,30 | 17,181 | 19,37 | 19,174 |
| 20,20 | 20,191 | 21,87 | 21,124 | 23,66 | 23,145 | 24,59 | 24,152 | 26,103 | 26,108 |

Table: 1 Some Points in $E_{211}(0,-4)$

Using (2,2) as Generator point ,we get the multiple point scalar multiplication of generator from 0 to Infinite Limit. Let $1 G = (2 , 2)$ to generate $2 G$ we perform scalar Multiplications $G + G = 2G$ and using the formulae

$$x_3 = \lambda^2 - x_1 - x_2, y_3 = \lambda (x_1 x_3 - y_1) \text{ and } \lambda = (y_2 - y_1) / (x_2 - x_1) \text{ if } P \neq Q$$

$$\lambda = 3x_1^2 + a / 2y_1 \text{ if } P = Q$$

$$= 12 / 4 = 3,$$

$$x_3 = 9 - 2 - 2 = 5,$$

$$y_3 = 3(2-5) - 2 = 3(-3) - 2 = -11 \text{ mod } 211 = 200$$

ANS:5,200

3.4.Generation of Points using Scalar Multiplication

To get $3G$ we add $G(2,2)$ to $2G(5,200)$ we get different P values as shown below

| | | | | | | | | | |
|-------|--------|--------|---------|---------|---------|---------|---------|---------|--------|
| 2,2 | 5,200 | 129,56 | 159,114 | 153,108 | 125,152 | 179,199 | 174,163 | 111,145 | 75,90 |
| 168,6 | 155,96 | 21,87 | 201,85 | 28,2 | 181,209 | 150,85 | 198,139 | 161,142 | 54,138 |
| 27,30 | 84,210 | 87,50 | 192,201 | 69,20 | 51,136 | 182,100 | 64,194 | 29,139 | 70,200 |

Table:2 The Scalar Multiplication points for $E_{211}(0,-4)$ at $G(2,2)$

This procedure is used for generating different elliptic curves. In fact we can use the same elliptic curve for all classes by changing the generator value. For the above points we can generate 240 different sets each containing 240 points for encryption.

4.Elliptic Curve Cryptography for defense messaging system

4.1.System Overview

Multilevel Access Control in Manet for a defense messaging system is useful for military organizations which have a hierarchical structure. For example in the Indian Military System the following hierarchy exists.

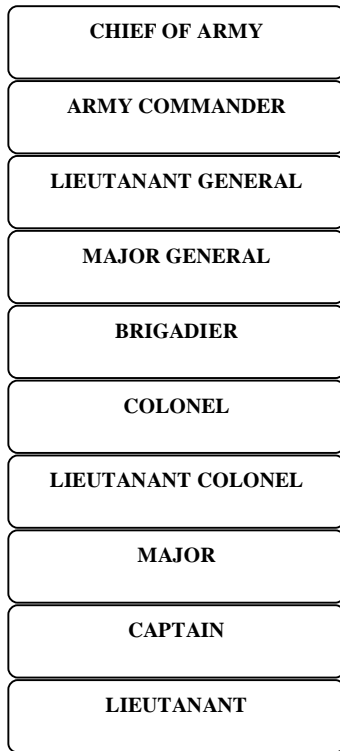


Figure 2 : Military hierarchy

In such a type of system, messages sent to a lower class should be known to the active members of lower class and also to all active members of the higher class. It is not only essential to maintain the access control but the data should be hidden as well. Elliptic curve cryptography technique is used .

There are many messages to be sent to different parties. The server inserts new data streams according to the classification . The messages are encrypted using ECC according to the access allowed for each user and ,the data is sent . Consider the following set of messages.

| Class | Category of Data Streams | | | |
|-------------|--------------------------|----------------|-----------------|-----------------|
| | Confiden tial | Field Messages | Terror Messages | Climate Warning |
| Troops | x | √ | x | x |
| Air Wing | x | x | x | √ |
| NSGS | x | x | √ | x |
| Lieutenants | √ | √ | x | x |

Table:3 Example Showing Message classifications

All the users of defense messaging system need to register themselves and get authenticated by the server. Once the registration process is over the user when joins receives the message he is entitled to receive .Only authenticated users are able to view the message content as the message remains unintelligible to people who do not belong to that elliptic curve. Different Elliptic curves identify different class of users .the servers contribution and the users contribution are used for finding the Group keys.

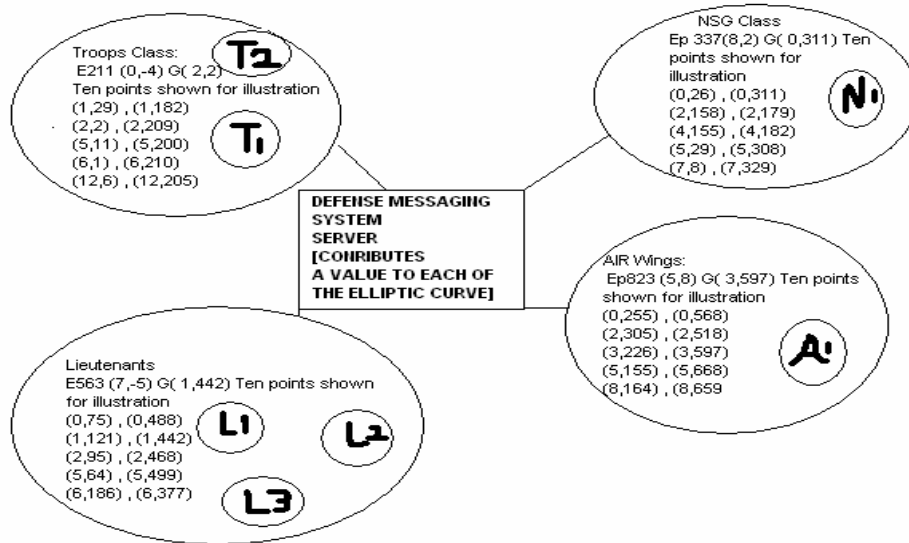


Figure 3 :System Overview

L1,L2,L3 belongs to lieutenant group ..N1 is National Security Guard ,T1,T2 are troops and A1 is a Air wing Officer.

4.2 The Proposed Scheme

The following methodology is used for developing a multilevel access control solution in a Manet for a defense messaging system using Elliptic Curve Cryptography.

Step1: The server or the central authority creates different classes with different service requirements. The system is dynamic and any hierarchy can be created or modified. The classes are arranged as a Tree based structure. A class can have any number of descendent classes at the next level and a class can have any number of users .

Step2 : Each class maintains a list of its ancestor and descendent classes.

Step 3: Each class is associated with an elliptic curve $E : y^2 = (x^3 + ax + b) \text{ mod } p$ over Z_p and $G = E(Z_p)$ is a Generating point.

Step 4: Users generate a random number as their contribution value and join their classes. The Shared secret key for each class is calculated whenever user joins or leaves .

Step 5: One user in each class is designated as the class controller . The user joining last is usually given the role of the class controller .

Step 6: The class controller of each class initiates the process of changing the group key Whenever users join or leave . Hence forward secrecy and backward secrecy is strictly maintained and only active members of the particular class will be able to view the message transfer of that particular class.

Step 7: Whenever the class controller leaves , the user who joined the class before the class controller now becomes the class controller. This is based on the assumption that a person who joins late will leave late which is many times true. The group key is changed in this case also so that the old class controller should not be able to decrypt the message.

Step 8: The group key is sent to all the ancestor class controller nodes by the class controller.

Step 9: The messages are also sent to the ancestor class controllers.

$$\begin{aligned} &= ((13525*91)\text{mod } 241) G \\ &= 229 G \\ &= (155,115) \end{aligned}$$

Troop User₂ Joins the Group

User Id: TUser2 Private Key (nC) = 82910
Public key (C) = $g^{nC} = (nC \text{ mod } p) G = (82910 \text{ mod } 241) G$
 $= 6 G = (125,152)$

Finding the Group key after the Second Troop User joined the Group

The new TUser2 act as a Group controller.

TUser2 computes g^{nBnC} , g^{nAnC}

$$g^{nA} = (206,121) \text{ yields } 91$$

$$g^{nAnC} = (91*82910 \text{ mod } 241) G = 64 G = (147,97)$$

$$g^{nB} = (29,139) \text{ yields } 29$$

$$g^{nBnC} = (29*82910 \text{ mod } 241) G = 174 G = (131,84)$$

Sends the g^{nBnC} Value to Server and g^{nAnC} Value to TUser1.

Finding the Group key after three users joined the group

Server Calculates the Group key

server will get g^{nBnC} from TUser₂ (GC) i.e. (131,84) yields 174

$$\begin{aligned} \text{Shared key} &= g^{nAnBnC} \\ &= ((47568*174)\text{mod } 241) G \\ &= 169 G \\ &= (120,31) \end{aligned}$$

TUser1 Calculates the Group key

TUser₁ will get g^{nAnC} from TUser₂ (GC) i.e. (147,97) yields 64

$$\begin{aligned} \text{Shared key} &= g^{nAnBnC} \\ &= ((13525*64)\text{mod } 241) G \\ &= 169 G \\ &= (120,31) \end{aligned}$$

TUser2 Calculates the Group key

g^{nAnB} i.e. (155,115) yields 229

$$\begin{aligned} \text{Shared key} &= g^{nAnBnC} \\ &= ((82910*229)\text{mod } 241) G \\ &= 169 G \\ &= (120,31) \end{aligned}$$

User Leave from the Group

Let the TUser2 be leave. Then the user sends message to all users that it is leaving. All the users remove the leaving user from the user list. The group controller changes its key value and computes the new group key.

Group controller New Private Key = 43297.

The group controller recalculates the following values:

$$g^{nAnB} = (155,115) \text{ yields } 229$$

Sends the g^{nB} Value to Server, g^{nA} Value to TUser1.

Using the shares the Group keys are calculated

Message Encryption

Message: Enter nestFire

Random Number K_A : 202

Cipher Text $P_c = (K_A G, PM + K_A S_K)$

$$\begin{aligned}K_A G &= 202 G \\ &= 202 \bmod 241 G \\ &= (50,57)\end{aligned}$$

Cipher text: 50:57:

1:182:164:197:203:180:1:182:172:50:136:11:164:197:
1:182:160:8:203:180:114:113:185:199:172:50:1:182:

PM+K_AS_K

PM

e→101 (1,182)
n→110 (164,197)
t→116 (203,180)
e→101 (1,182)
r→114 (172,50)
(Space)→32 (136,11)
n→110 (164,197)
e→101 (1,182)
s→115 (160,8)
t→116 (203,180)
f→102 (114,113)
i→105 (185,199)
r→114 (172,50)
e→101 (1,182)

$$\begin{aligned}K_A S_K &= (202 * 18) \bmod 241 \\ &= 21\end{aligned}$$

PM+K_AS_K

$$\begin{aligned}e &= (101+21) \bmod 241 \\ &= 122 G \\ &= (156,10) \\ n &= (110+21) \bmod 241 \\ &= 131 G \\ &= (164,14) \\ t &= (116+21) \bmod 241 \\ &= 137 G \\ &= (163,161) \\ e &= (101+21) \bmod 241 \\ &= 122 G \\ &= (156,10) \\ r &= (114+21) \bmod 241 \\ &= 135 G \\ &= (77,145) \\ &= (32+21) \bmod 241 \text{ (for space)} \\ &= 53 G \\ &= (99,180) \\ n &= (110+21) \bmod 241 \\ &= 131 G \\ &= (164,14) \\ e &= (101+21) \bmod 241 \\ &= 122 G \\ &= (156,10) \\ s &= (115+21) \bmod 241 \\ &= 136 G \\ &= (185,12) \\ t &= (116+21) \bmod 241 \\ &= 137 G = (163,161)\end{aligned}$$

$$\begin{aligned}
 f &= (102+21) \bmod 241 \\
 &= 123 \text{ G} \\
 &= (104,190) \\
 i &= (105+21) \bmod 241 \\
 &= 126 \text{ G} \\
 &= (160,203) \\
 r &= (114+21) \bmod 241 \\
 &= 135 \text{ G} \\
 &= (77,145) \\
 e &= (101+21) \bmod 241 \\
 &= 122 \text{ G} \\
 &= (156,10)
 \end{aligned}$$

Message Decryption

The Group Key (S_k) is (198,139)
 Random no. Chosen can be found by the $K_A G$ Value (50,57).

From (50,57) we may trace the value K_A as 202.

$$(1) PM + K_A S_K = (156,10)$$

$$\begin{aligned}
 PM &= (156,10) - K_A S_K \\
 &= (156,10) - 21 \\
 &= 122 - 21 \\
 &= 101 \text{ G} \quad \Rightarrow (1,182) \rightarrow e
 \end{aligned}$$

$$(2) PM + K_A S_K = (164,14)$$

$$\begin{aligned}
 PM &= (164,14) - K_A S_K \\
 &= (164,14) - 21 \\
 &= 131 - 21 \\
 &= 110 \text{ G} \quad \Rightarrow (164,197) \rightarrow n
 \end{aligned}$$

$$(3) PM + K_A S_K = (163,161)$$

$$\begin{aligned}
 PM &= (163,161) - K_A S_K \\
 &= (163,161) - 21 \\
 &= 137 - 21 \\
 &= 116 \text{ G} \quad \Rightarrow (203,180) \rightarrow t
 \end{aligned}$$

$$(4) PM + K_A S_K = (156,10)$$

$$\begin{aligned}
 PM &= (156,10) - K_A S_K \\
 &= (156,10) - 21 \\
 &= 122 - 21 \\
 &= 101 \text{ G} \quad \Rightarrow (1,182) \rightarrow e
 \end{aligned}$$

$$(5) PM + K_A S_K = (77,145)$$

$$\begin{aligned}
 PM &= (77,145) - K_A S_K \\
 &= (77,145) - 21 \\
 &= 135 - 21 \\
 &= 114 \text{ G} \quad \Rightarrow (172,50) \rightarrow r
 \end{aligned}$$

- (6) $PM+K_A S_K = (99,180)$
 $PM = (99,180) - K_A S_K$
 $= (99,180) - 21$
 $= 53 - 21$
 $= 32 G \Rightarrow (136,11) \rightarrow (\text{Space})$
- (7) $PM+K_A S_K = (164,14)$
 $PM = (164,14) - K_A S_K$
 $= (164,14) - 21$
 $= 131 - 21$
 $= 110 G \Rightarrow (164,197) \rightarrow n$
- (8) $PM+K_A S_K = (156,10)$
 $PM = (156,10) - K_A S_K$
 $= (156,10) - 21$
 $= 122 - 21$
 $= 101 G \Rightarrow (1,182) \rightarrow e$
- (9) $PM+K_A S_K = (185,12)$
 $PM = (185,12) - K_A S_K$
 $= (185,12) - 21$
 $= 136 - 21$
 $= 115 G \Rightarrow (160,8) \rightarrow s$
- (10) $PM+K_A S_K = (163,161)$
 $PM = (163,161) - K_A S_K$
 $= (163,161) - 21$
 $= 137 - 21$
 $= 116 G \Rightarrow (203,180) \rightarrow t$
- (11) $PM+K_A S_K = (104,190)$
 $PM = (104,190) - K_A S_K$
 $= (104,190) - 21$
 $= 123 - 21$
 $= 102 G \Rightarrow (114,113) \rightarrow f$
- (12) $PM+K_A S_K = (160,203)$
 $PM = (160,203) - K_A S_K$
 $= (160,203) - 21$
 $= 126 - 21$
 $= 105 G \Rightarrow (185,199) \rightarrow i$
- (13) $PM+K_A S_K = (77,145)$
 $PM = (77,145) - K_A S_K$
 $= (77,145) - 21$
 $= 135 - 21$
 $= 114 G \Rightarrow (172,50) \rightarrow r$
- (14) $PM+K_A S_K = (156,10)$
 $PM = (156,10) - K_A S_K$
 $= (156,10) - 21$
 $= 122 - 21$
 $= 101 G \Rightarrow (1,182) \rightarrow e$

4.4.Elliptic Curves used

Troops Class:

$$y^2 = x^3 - 4 \text{ mod } 211 \text{ at } G(2,2)$$

NSG Class: (National Security Guards)

$$y^2 = x^3 + 8x - 2 \text{ mod } 337 \text{ at } G(0,311)$$

Class Join

Enter the Class Name:

Select the Services You Need:

- Confidential
- FieldMessage
- TerroristInformation
- ClimateCondition

Figure:5 Class Join

Class Delete

Select the Class to Delete

Figure:6 Class Delete

User Join

Enter the User Name:

Enter the Password:

Re enter the Password

Enter the Class you want to join:

Figure:7 User Join

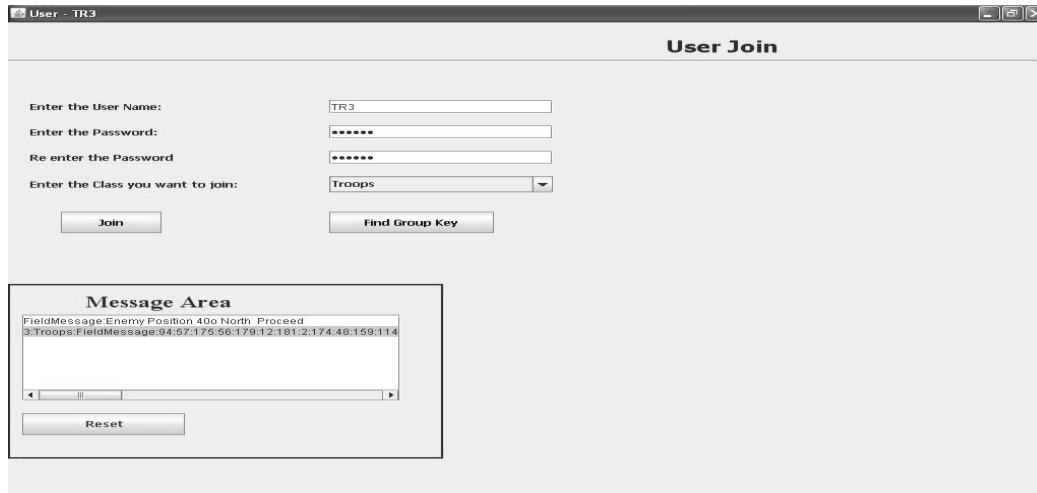


Figure :8 Message Received Encrypted and decrypted

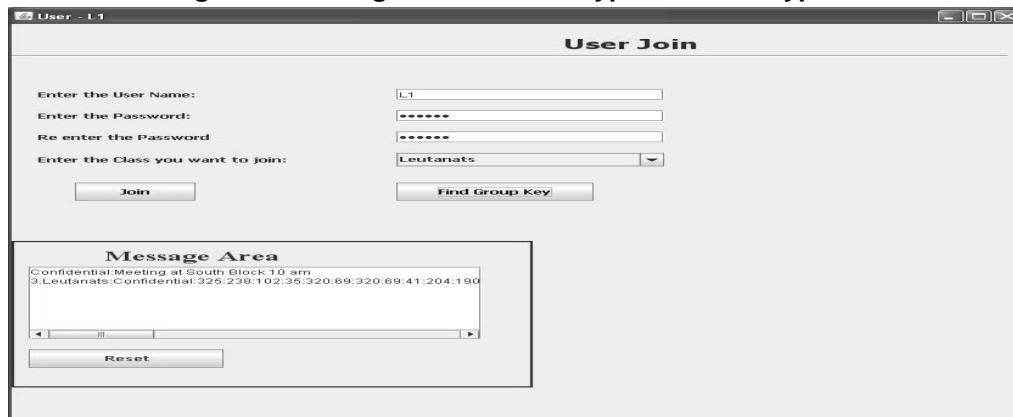


Figure :9 Message Received Encrypted And Decrypted by lieutenant class

6. PERFORMANCE ANALYSIS:

6.1. Security :

The Security of ECC is due to the discrete logarithm problem over the points on the elliptic curve. Cryptanalysis involves determining x given Q and P where P is a point on the elliptic curve and $Q = x P$ that is P added to itself x times. The best known algorithm to break the elliptic curve points is the pollard – rho algorithm which is a fully exponential algorithm and difficult o solve.. Forward and Backward secrecy are maintained as each session is considered as a separate session. In this section we discuss some attacks and prove that our scheme is secure and feasible. Consider the Figure 10

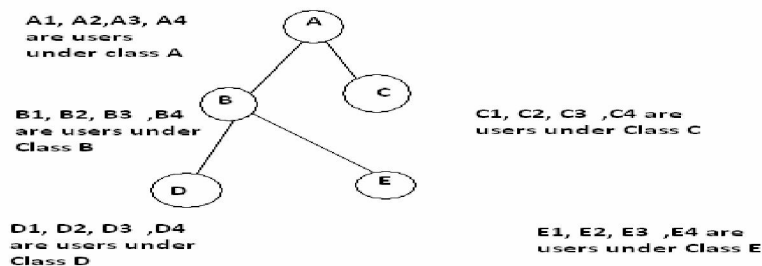


Figure :10 Illustration for security

Attack 1: Contrary Attacks

Assuming that E1 (lower privileged user) wants to crack the secret key of B1 (Higher Privileged User). It is not possible to decrypt any messages as the hierarchy does not provide secret keys to descendent classes. Without knowing the secret key it is impossible to see the message.

Attack 2: Interior Collecting Attacks

If a lower level User has many ancestors and if it negotiates with one parent also by knowing the key as there is no relation parameter between any of the ancestor nodes it is not possible to derive the key.

Attack 3: Exterior Collecting Attack

If an attacker is outside the system, it means no idea about what elliptic curve or generator point is being used is known and hence more difficult to attack.

Attack 4: Collaborative Attacks

We assume that if there is a higher privileged user belonging to class B as in figure 10 and there are two descendant classes D and E. Users of D and E cannot perform a collaborative attack as the secret key of any class is calculated only from the contribution of the respective users of the class. If they compromise one user belonging to the higher privileged group to know the key also, there is no communication possible as the class controllers who control the keying process also change dynamically.

Attack 5: Sibling Attacks

Classes who have same parent also cannot crack the key of a sibling class due to the absence of any related parameters among them.

To maintain the secure structure the following things are necessary.

1. The immediate parents should be loyal and the descendant list should be updated.
2. The Class Controller of a Leaving /Changing Class from the tree hierarchy should update their ancestor list.
3. Recalculation of shared secret key by the leaving /changing class should be done by selecting a random value for finding a new group key.

The third point is very important and the execution of this prevents disloyal ancestors also from finding the key of a left descendant class even if it no longer comes under their control. The most important feature is that only the key is being transferred and only the authorized entities have idea about which elliptic curve is used or generator points that are used. This is a very big advantage as even though an adversary comes across the key pair he may not know the elliptic curve and the generator. Even users inside the system may decrypt the information but they may not be aware about the mechanism that takes place.

6.1.1 Enhanced Security Framework:

The keys are always transmitted as plain text but we have justified that even with the key any attacker will not be able to first of all receive the messages and even if they somehow receive the message and the key, they will not be able to decrypt the information as they are unaware about the elliptic curve. If security needs to be enhanced the following framework can be included

1. All first level ancestor – descendant pairs use Diffie- Hell man key exchange and generate a key which can be used for encrypting and decrypting the actual key.
2. This key can again be combined with other descendant classes. For e.g., in our sample hierarchy first we perform key exchanges and get the shared secret keys SSK BA, SSK CA and then again use classes D and E and generate SSK DBA, SSK EBA we can use the resultant keys for encrypting and decrypting the original keys.
4. The above scheme prevents repeated computations

6.2. Memory Cost:

Using ECC approach consumes very less memory when compared to RSA and DES. ECC based approach takes very less memory even the members get increased.

6.3.Communication Costs:

Using other schemes consumes more bandwidth. The Communication and computation of tree based ECDH depends on trees height, balance of key tree, location of joining tree, and leaving nodes. But our approach depends on the number of member in the subgroup, number of Group Controller, and height of tree. So the amount spend on communication is very much less when compared to CRTDH and RSA based scheme.

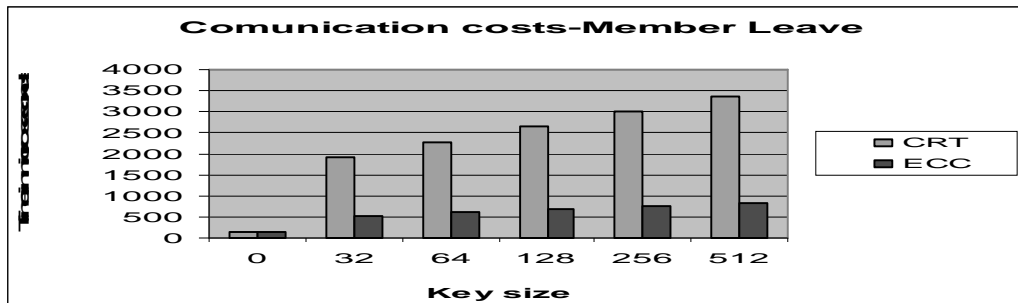


Figure 11.Communication Cost For Member Leave

Consider (Figure.10& 11) there were 256 members in a group our approach consumes only 29% of Bandwidth when compare to CRTDH and RSADH. So our approach consumes low Bandwidth.

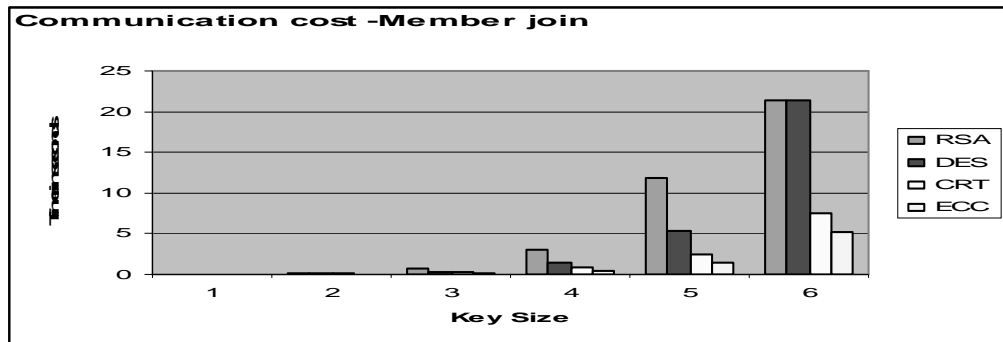


Figure 12. Communication cost for Member Join

For member leave operation also our approach takes less time as the key size for ecc is small compared to other approaches.

6.4. Computation Costs:

The Computational costs depend on the Number of exponentiations. CRTDH have high computation costs as it depends on the number of members and group size respectively. The cost increases as the members and group size increases. But our approach spends a little on this computation. The number of bits for encryption is very less compared to other keys .Moreover each user need not store any data key values.

7. CONCLUSION AND FUTURE WORK:

MANETs are much more vulnerable to attack than wired network. This is because of the following reasons :

Open Medium - Eavesdropping is more easier than in wired network.

Dynamically Changing Network Topology - Mobile Nodes comes and goes from the network, thereby allowing any malicious node to join the network without being detected.

Cooperative Algorithms - The routing algorithm of MANETs requires mutual trust between nodes which violates the principles of Network Security.

Lack of Centralized Monitoring - Absence of any centralized infrastructure prohibits any monitoring agent in the system.

Lack of Clear Line of Defense – There is a lack of clear line of defense - attack prevention may not suffice.

We have implemented on Multilevel Access Control in a Manet for Defense Messaging System using Elliptic curve cryptography. Use of Elliptic curve ensures that data is protected and intruders cannot guess the message. Moreover a single message sent will reach all the classes which are higher in the hierarchy. The system satisfies the user dynamics and class dynamic. We have successfully implemented by selecting different elliptic curves. A single elliptic curve can be used and by changing the generator points and we can perform different encryption. The group keys are found by the server and forward and backward secrecy is maintained here. The user level, DataStream level and class level hierarchies are taken care. As future implementation new methods for improving parameters can be done..

8.REFERENCES

- [1] S. Akl and P. Taylor. "Cryptographic solution to a problem of access control in a hierarchy". ACM Transactions on Computer Systems, 1(3):239{248,September 1983.
- [2] William Stallings," Cryptography and network security Principles and Practices",Third Edition, Pearson education.2001
- [3] M. Atallah, K. Frikken, and M. Blanton," Dynamic And Efficient Key Management For Access Hierarchies", CERIAS Tech Report 2006-02,Center for Education and Research in,Information Assurance and Security,Purdue University,
- [4] Jason Crampton,"Cryptographically-Enforced Hierarchical Access Control with Multiple Keys", Journal of Logic and Algebraic Programming April 1, 2009.
- [5] K. Kumar J.Nafeesa Begum, Dr.V.Sumathy, (2009), "A Novel Approach towards cost Effective Region Based Group Key Agreement Protocol for Ad Hoc Networks" Computational Intelligence, Communication Systems and Networks ,2009 CICSYN,09 ,July 23-25 2009 published in IEEE Explore.
- [6] K.Kumar, J.Nafeesa Begum ,Dr.V.Sumathy , ,(2009), " Efficient Region-Based Group Key Agreement Protocols for Ad Hoc Networks using Elliptic Curve Cryptography",IEEE International Advance Computing Conference(IACC-2009), Thapar University, Patiala March 6-7 . published in IEEE Explore.

- [7] Tim Bauge, White paper on "Ad hoc networking in military scenarios", Thales Research and Technology (UK) Limited , May 2004
- [8] S. J. MacKinnon, P. D. Taylor, H. Meijer, and S. G. Akl." An optimal algorithm for assigning cryptographic keys to control access in a hierarchy". *IEEE Transactions on Computers*, 34(9):797.802, Sept. 1985.
- [9]S. Chen, Y.-F. Chung, and C.-S. Tian. "A novel key management scheme for dynamic access control in a user hierarchy", In *COMPSAC*, pages 396.397, Sept. 2004.
- [10] I. Ray, I. Ray, and N. Narasimhamurthi. A cryptographic solution to implement access control in a hierarchy and more. In *SACMAT '02: Proceedings of the seventh ACM symposium on Access control models and technologies*, pages 65.73. ACM Press, 2002.
- [11] R. S. Sandhu. "Cryptographic implementation of a tree hierarchy for access control". *Information Processing Letter*, 27(2):95.98, Feb. 1988.
- [12] G. C. Chick and S. E. Tavares."Flexible access control with master keys", *Proceedings on Advances in Cryptology: CRYPTO '89, LNCS*, 435:316.322, 1989.
- [13] M. L. Das, A. Saxena, V. P. Gulati, and D. B. Phatak." Hierarchical key management scheme using polynomial interpolation". *SIGOPS Operating Systems Review*, 39(1):40.47, Jan. 2005.
- [14] L. Harn and H. Y. Lin." A cryptographic key generation scheme for multilevel data security". *Computers and Security*, 9(6):539.546, Oct. 1990.
- [15] V. R. L. Shen and T.-S. Chen. ,"A novel key management scheme based on discrete logarithms and polynomial interpolations". *Computers and Security*, 21(2):164.171, Mar. 2002.
- [16] M.-S. Hwang, C.-H. Liu, and J.-W. Lo,". An efficient key assignment for access control in large partially ordered hierarchy". *Journal of Systems and Software*, Feb. 2004.
- [17] C. H. Lin. "Dynamic key management scheme for access control in a hierarchy. *Computer Communications*,20(15):1381.1385, Dec. 1997.
- [18] S. Zhong. A practical key management scheme for access control in a user hierarchy". *Computers and Security*, 21(8):750.759, Nov. 2002.
- [19] X. Zou, B. Ramamurthy, and S. Magliveras. "Chinese remainder theorem based hierarchical access controlfor secure group communications". *Lecture Notes in Computer Science (LNCS)*, 2229:381.385, Nov. 2001.
- [20] X. Zou, B. Ramamurthy, and S. S. Magliveras, editors." *Secure Group Communications over Data Networks*",Springer, New York, NY, USA, ISBN: 0-387-22970-1, Oct. 2004.

FPGA Prototype of Robust Image Watermarking For JPEG 2000 With Dual Detection

Pankaj U.Lande

Dept. Instrumentation Science,
University of Pune,
Pune

pul@usic.unipune.ernet.in

Sanjay N. Talbar

S.G.G.S. Institute of Engineering and Technology,
Nanded.

sntalbar@yahoo.com

G.N. Shinde

Indira Gandhi College CIDCO,
Nanded.

shindegn@yahoo.co.in

Abstract

This paper presents a novel robust invisible watermarking method for still images. The scheme is implemented on hardware, and it can be incorporated with the lossless JPEG2000 compression standard. We have implemented Cohen-Daubechies-Favreau (CDF) 5/3 wavelet filters with lifting scheme which requires less hardware and they are also the basis of lossless JPEG2000. Its modular structure is well suitable for hardware implementation and it is more efficient use of power and chip area. The objective of the hardware assisted watermarking is to achieve low power usage, real-time performance, robust and ease of integration with existing consumer electronic devices such as scanners, cameras and handy camcorders. The proposed scheme of watermarking is tested with StirMark software which is a one of the benchmarking software for watermarking scheme. The experimental result shows that the proposed scheme of watermarking is robust against most of the geometric attacks such as scaling and rotation. We have proposed a dual detection technique for watermark detection which is a novelty of our algorithm.

Keywords: CDF 5/3 wavelet, FPGA, watermarking

1. INTRODUCTION

The recent proliferation and success of the internet, together with the availability of relatively inexpensive digital recording and storage devices has created an environment in which it becomes very easy to obtain, replicate and distribute digital content without any loss in quality. This growth of applications in the past decade gave rise to the new set of problems like *digital piracy*: illegal copying, use, and distribution of copyrighted digital data. This has become a great concern to the multimedia content such as music, video and image to the publishing industries, because technologies or techniques to protect intellectual property rights for digital media and to prevent unauthorized copying did not exist. Exactly identical copies of digital information, be it images, text or audio, can be produced and distributed easily. In such a scenario, who is the artist and who the plagiarist? It's impossible to tell or was, until now. Digital right management (DRM) is a collection of technologies and a technique that enables the licensing of digital information including the multimedia content such as image, video and music. DRM consist of two prominent technologies those are encryption and watermarking. Encryption technologies can be used to prevent unauthorized access to digital content. However, encryption has its limitations in protecting intellectual property rights, because once digital content is decrypted, there is nothing to prevent an authorized user from illegally replicating it [1][2].

Digital watermarking is a process in which an informed signal (watermark) is incorporated in multimedia content such as images to protect the owner's copyright over that content. The watermark can be later be extracted from a suspected image and be verified in order to identify the copyright owner. Watermarking technique for paper manufacturing have been in use since the middle ages, same concept was adopted by digital world and extended this concept for digital images, video and music.[3]. A watermarking scheme consists of three parts: the watermark, the encoder, and the decoder. The watermarking algorithm incorporates the watermark in the object, whereas the verification algorithm authenticates the object by determining the presence of the watermark and its actual data bits [4].

Watermarking techniques can be divided into various categories in numerous ways [5]. In the case of still digital images, there are three primary methods for insertion and extraction of a watermark. These are spatial domain, transform domain and color space methods. The spatial domain method [6] involves an algorithm that directly operates on the pixel values of the host image. In the transform domain method the pixel values are transformed into another domain by applying appropriate transform technique like discrete cosine transform (DCT) [7][9][10], discrete wavelet transform(DWT)[8][11] and Hadamard transform[12]. A watermark is then embedded by modifying these coefficients. However it is observed that spatial domain watermarks are weaker than frequency domain ones [13][14]. A DCT based watermarking algorithm has been described in many literatures; however DWT based watermarking algorithms are more effective for several reasons [15].

Wavelet is a small wave whose energy is concentrated in time and still possesses the periodic characteristics. An arbitrary signal can be analyzed in terms of scaling and translation of a single mother wavelet function. Properties of wavelet allows both time and frequency analysis of signals simultaneously. They offer excellent space-frequency localization of salient image features such as textures and edges. DWT can analyze the data in different scales and resolutions this principal is called as multi-resolution analysis [16]. DWT decomposed the signal into lower and higher frequency signal components. The high-frequency content of an image corresponds to a large coefficient in the detail sub-band. Hence, watermark encoders operating in the wavelet domain can easily locate the high-frequency features of an image and embed most of the watermark energy. Such a method of embedding results in an implicit visual masking of the watermark, because the ability of human visual system (HVS) to detect high frequency signals is limited [17]. It is also a basis of a compression standards like JPEG2000 [18] and MPEG-4[19].

1.1 Related Work

Software approach for image watermarking have been proposed in many literatures; but hardware implementations has few advantages over software approach such as

1. It gives optimized specific design which is a small, fast, and potentially cheap watermarking unit.
2. It is most suitable for real-time applications, where the computation time is deterministic and short.
3. Hardware based watermarking unit can be easily integrated with digital cameras and scanners, graphics processing units etc.
4. Hardware watermarking unit consumes lesser power than software, which requires a general purpose processor so that they are ideal for battery operated applications.
5. The cost is low compared to that software used explicitly for watermarking; this is because a hardware based watermarking unit can be monolithically built on a single unified system in the context of system-on-chip (SoC) technology.

6. The hardware can be implemented as a soft core expressed in the structural hardware description language like VHDL and Verilog. The soft core can be modified as algorithm changes and can be resynthesized into new silicon technology.

A hardware based watermarking is presented in few literatures illustrated bellow.

Seo and Kim [20] presented a field programmable gate array (FPGA) based implementation of blind and invisible watermarking on Altera FPGA. This watermarking algorithm was presented in DCT domain and the DC coefficients are replaced by watermark. The two dimensional DCT was calculated for one or more than one bit planes and the DC coefficients are replaced in such a way that it will be imperceptible to human eyes. The watermarking algorithm was integrated with JPEG2000 encoder and it operates on 66MHz.

An FPGA based invisible robust spatial domain watermarking is described in [21].the watermark insertion is carried out by replacing original image pixel value by watermark encoding function. The watermark is generated through user key. The watermarking scheme is evaluated by standard benchmark like StirMark software. The original image is required for watermark detection. The algorithm was implemented on XCV50-BG256-6 device from Xilinx and operated on 50.398MHz.

An FPGA prototype of Biometric based watermarking is described in [22]. The algorithm work for both gray and color image and the biometric image is selected as watermark. The original image is divided in 8x8 blocks and DCT is calculated for each block. The biometric image (watermark) is divided into blocks and embedded into perceptually significant region of cover image. This approach makes the watermark robust against the common signal processing attacks. Original image is required watermark detection. The prototype was modeled using VHDL and implemented on XC2V500-6FG256 device from Xilinx.

Saraju P. Mohanty et. al. [23] proposed a novel algorithm for encrypted watermarking based on block-wise DCT. The watermarking can work for gray scale image as well as color image. In the case of color image the cover image in RGB format is converted into YCbCr and the Y component is selected for watermarking. The image is divided in 8x8 blocks and DCT is calculated for each block. The encrypted watermark is embedded into transformed image by four different embedding factors. The embedding strength factor is chosen such that the image quality will not degrade. The watermark detection process requires original image. The block-wise DCT is computed for both image and the difference is calculated to detect a watermark. The extracted watermark is compared with original watermark to authenticate the suspected image.

Image adaptive watermarking and its hardware architecture is described in [12]. The proposed scheme of watermarking is invisible and robust against JPEG attacks. Cover image is divided in 8x8 blocks and DHT is calculated. PN sequence is generated through user key and embedded into DHT coefficients. The strength factor is calculated from quantization table for DHT domain. Watermark detection method is blind. The proposed method is robust against the common signal processing attacks like median filtering and noise addition. The algorithm was implemented on XC3SD1800A-4FGG676C and functional simulation was performed using Xilinx tools. The chip was tested using hardware co-simulation which was run at 33.3MHz.

In this paper we have presented a watermarking scheme using the CDF 5/3 wavelet filter which can be incorporated with JPEG2000 lossless image compression. Hardware architecture was implemented on FPGA. The proposed scheme is an invisible robust wavelet domain watermarking method. We have also proposed a dual watermark detection technique that is the watermark can be detected by blind and non blind method to verify the suspected image. The blind watermark detection can be used with the images from digital cameras where the original image is not present. The non-blind technique can be used with the digital scanners where the original image is present.

The scheme described in [24] is used to implement CDF 5/3 wavelet filters. The proposed architecture uses the lifting scheme technique and provides advantages that include small memory requirements, fixed-point arithmetic implementation, and a small number of arithmetic computations. The chip was modeled using Verilog and a function simulation was performed. This chip was tested using AccelDSP in hardware in the loop (HIL) arrangement. The proposed scheme is robust against several geometric attacks. We have tested our watermarking scheme using standard benchmark such as StirMark software.

2. PROPOSED WATERMARKING SCHEME

The proposed scheme is based on CDF 5/3 wavelet filters which is the basis of lossless JPEG2000 compression standard. The new still compression image standard, JPEG2000 has emerged with a number of significant features that would allow it to be used efficiently over a wide variety of images. The JPEG2000 standard exhibits a lot of features, the most significant being the possibility to define regions of interest in an image, the spatial and SNR scalability, the error resilience and the possibility of intellectual property rights protection. Interestingly enough, all these features are incorporated within a unified algorithm. This compression standard uses the Cohen-Daubechies-Favreau (CDF) 5/3 and CDF 9/7 DWT for lossless and lossy image compression respectively. Since JPEG2000 is the newest version of one of the most popular image formats and it includes the DWT, efficient VLSI implementations of DWT processors became more and more important.

We have used the lifting scheme described in[24].the advantages of using lifting scheme is that, the number of multiplications and additions compared to the filter-bank implementation are reduced resulting in more efficient use of power and chip area. The modular structure is well suitable for hardware implementation. The lifting scheme calculates the DWT using spatial domain analysis, and consists of a series of *Split*, *Predict* and *Update* steps. The split step separates odd and even samples, and predict step predicts values in the odd set where $\alpha = -0.5$ as the predict step coefficient. The Update step uses the new wavelet coefficients in the odd set to update the even set, where $\beta = 0.25$ as the update step coefficient. Lifting scheme is shown in figure 1 and Lifting operation for the CDF 5/3 synthesis filter is shown in Figure 1.

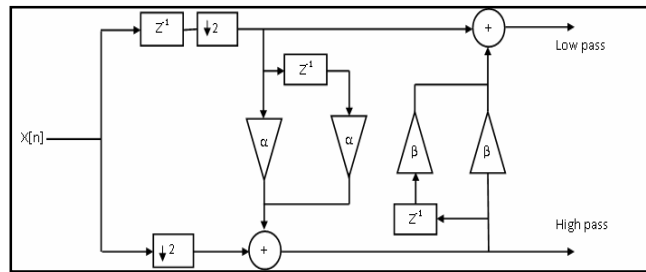


FIGURE 1: CDF 5/3 with lifting scheme

The watermarking algorithm embeds same multiple watermarks in cover image. The cover image I is divided into non-overlapping blocks of size $B \times B$. CDF 5/3 wavelet transform is calculated for block separately. A binary watermark is embedded into cover image using equation (1).

$$I_{W,N}(x, y) = I_N(x, y) + a \times W(x, y) \quad (1)$$

Where 'a' is gain

$I_{W,N}$ is a N^{th} block of watermarked image and W is a binary watermark logo. x and y are index numbers.

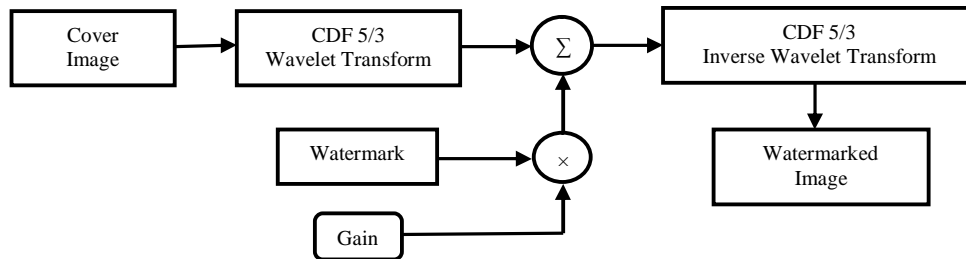


FIGURE 2: Watermarking Scheme

2.1 Hardware Architecture for Proposed Scheme

In this section, we discuss the hardware architecture for the scheme discussed in the previous section. The watermarking chip mainly consists of a block processing unit and control.

2.1.2 Block processing unit

The block processing unit considers the original image block as input. Image block is wavelet transformed and the watermark is embedded using equation (1). This unit consists of CDF5/3 wavelet filters and watermarking unit. To meet the real time constrain, we have used two filters in parallel to calculate forward and inverse transform. In order to calculate the 2D wavelet, these filters first calculate the coefficients first row-wise and then column-wise. The intermediate results are stored in the memory. Inverse wavelet is calculated in similar manner.

2.1.3 Watermarking unit

The watermarking unit consists of a multiplier and adder. The watermark is embedded using equation (1). Because a multiplier requires more hardware, only one multiplier is implemented. The wavelet transformed block is fed serially to the watermarking unit. The gain is multiplied by watermark and added to the wavelet transformed coefficients. The intermediate results are stored in the memory.

2.1.4 Control unit

The control unit generates the necessary control signals for the entire system during the watermarking process. The control unit generates four main signals and these signals are as follows

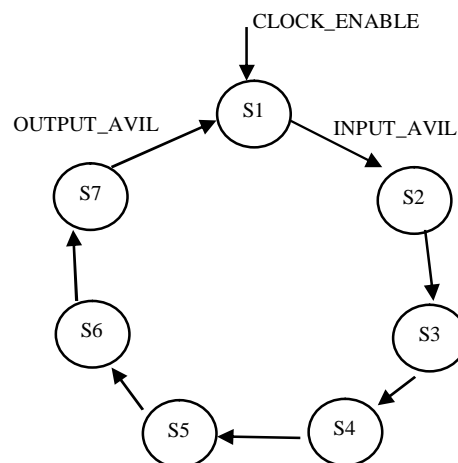


FIGURE 3: Control Unit as FSM

INPUT_AVIL: Image block is available at input.

OUTPUT_AVIL: Watermarked image block is available at output

CLOCK: Clock signal for chip

CLOCK_ENABLE: When clock enable is high chip is in an active mode for processing

This unit undergoes seven states in each state; the particular task is performed in each state and the finite state machine (FSM) begins to the next state. Figure (3) shows the state diagram of FSM.

S1: if the clock enable is high and INPUT_AVIL is high then read image block

S2: calculate DWT

S3: read the watermark

S4: multiply the watermark with 'a'

S5: embedded the watermark

S6: calculate inverse DWT

S7: generate OUTPUT_AVIL signal

2.1.5 Watermark detection

The watermark detection algorithm is implemented using MatLab. The watermark can be detected using two methods blind and non-blind. In non blind method original and watermarked image both are required to detect a watermark. The suspected image and original image are divided into BxB blocks, and DWT coefficients are calculated for both images. The watermark is recovered using equation (4).

$$W(i, j) = \begin{cases} 1 & \text{if } I_{WN}(x, y) - I_N(x, y) > \tau \\ 0 & \text{other wise} \end{cases} \quad (4)$$

τ represents threshold for blind detection

In the blind watermark detection method the binary logo image is considered as PN sequence and the correlation between the suspected image and watermark is calculated. The suspected image is divided into BxB blocks, and DWT coefficients are calculated. The correlation between encrypted watermark and wavelet transformed block is calculated using equation (5)

$$\gamma = \frac{\sum_m \sum_n (I_{WN}(x, y) - \bar{I}_{WN})(w(x, y) - \bar{w})}{\sqrt{\sum_m \sum_n (I_{WN}(x, y) - \bar{I}_{WN}) \sum_m \sum_n (w(x, y) - \bar{w})}} \quad (5)$$

If $\gamma > \rho$ then the watermark is detected. ρ is threshold for blind detection.

If $\gamma > \rho$ then the watermark is detected. ρ is threshold for blind detection.

3. EXPERIMENTAL RESULTS

3.1 Synthesis and Implementation

The chip was modeled using a Verilog and functional simulation was performed. The code was synthesized on Xilinx Spartan-3A technology on XC3SD1800A-4FGG676C device using the AccelDSP. The results are verified by hardware in the loop (HIL) configuration using AccelDSP. The HIL was run at 33.3 MHz clock frequency, and the samples were fed to the target device at a rate of 319.585 Ksps through a JTAG USB cable. The design utilizes 2 startup clock cycles and single clock cycles per function call. The device utilization summary is given in Table 1.

| Logic Utilization | Used | Available | Utilization % |
|----------------------------|------|-----------|---------------|
| Number of Slices | 628 | 16640 | 3.77 % |
| Number of Slice Flip Flops | 290 | 33280 | 0.87 % |
| Number of 4 input LUTs | 1077 | 33280 | 3.23 % |
| Number of bonded IOBs | 293 | 309 | 94.82 % |
| Number of GCLKs | 1 | 24 | 4.16 % |

TABLE 1: Device Utilization Summary

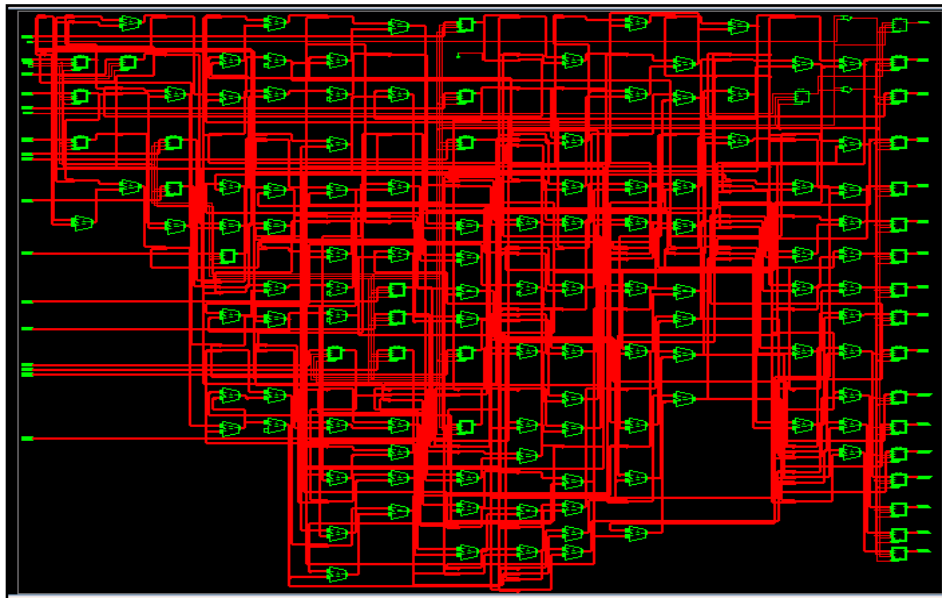


FIGURE 4: RTL of Watermarking Chip

3.2 Image Quality Measures

In [25] Kutter and Petitcolas have discussed various parameters to estimate any watermarking scheme. For fair benchmarking and performance evaluation, the visual degradation due to embedding is an important issue. Most distortion measure (quality metrics) used in visual information processing belongs to a group of difference distortion measures. The watermark images are acceptable to the human visual system if the distortion introduced due to watermarking is less.

The various performance evaluations metrics such as PSNR (db), Image Fidelity (IF), Normalized cross correlation, correlation quality etc. are calculated. Results for few popular images are given in Table 2.

| Quality Measures | Lena | Mandrill | Woman |
|------------------------------|---------|----------|---------|
| Mean square error | 6.80 | 6.74 | 6.80 |
| PSNR | 39.79 | 39.84 | 39.80 |
| Normalized cross correlation | 1 | 1 | 1 |
| Average Difference | -0.8135 | -0.8055 | -0.8146 |
| Structural content | 0.98 | 0.99 | 0.98 |
| Maximum difference | 3 | 3 | 3 |
| Normalized absolute error | 0.031 | 0.017 | 0.017 |
| Image Fidelity | 1 | 1 | 1 |
| correlation quality | 1 | 1 | 1 |

TABLE 2. Image Quality Measures



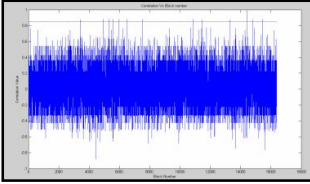


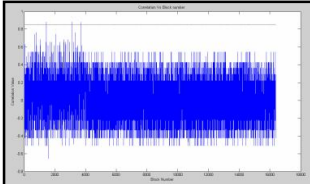
3.3 Performance Evaluations on Various Attacks

In this section, we evaluate the performance of the watermarking algorithm against various attacks using standard benchmark software. The StirMark software includes several attacks such as JPEG compression geometric transformation, noise addition etc. The geometric attacks includes rotation, cropping, scaling and geometric transformation with medium compression. Some of the results of these evaluations for blind and non-blind detection are summarized in Table 3. For blind detection threshold is $\rho=0.85$. These results indicate that the proposed watermarking scheme is robust against the geometric attacks.

The proposed scheme of watermarking embeds multiple watermarks in cover image. The objective was, at least a single watermark will survive after attacks. In detection algorithm all the watermarks are detected, and the watermark which is having highest correlation with the original watermark is treated as the recovered watermark. We have also proposed a dual watermark detection technique. The watermark can be detected by blind or non-blind method and both detection techniques can be used to verify suspected image. Scheme implements several watermark in the cover image, due to which scheme is robust against various geometric attacks.

4. Conclusion

In this paper, we proposed a novel invisible image watermarking algorithm and developed efficient the hardware architecture which can be used with JPEG2000. The watermarking scheme utilizes minimal hardware resources as it can be seen from the device utilization summary table. Because of the lifting scheme is used in CDF 5/3 filters it requires minimum hardware and it requires less clock cycles. The experimental results showed that the proposed scheme of watermarking scheme is imperceptible and robust against geometric attacks. The proposed algorithm outperforms than the presented algorithm in [12][21]. This was achieved because of space and frequency localizing property that is the characteristics of the discrete wavelet transform. In the future we want to develop a image adaptive watermarking hardware using fuzzy logic or neural network.

| Attacks | Non – Blind Detection | Blind Detection |
|---|---|--|
|  AFFINE_2 |  |  Max corr value =0.86 |
|  CROP_25 |  |  Max corr value =0.87 |

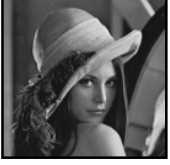

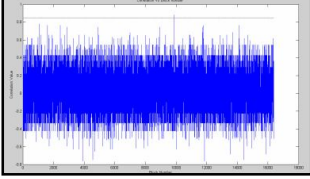


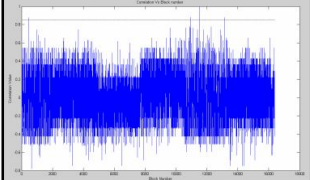


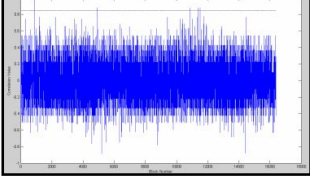


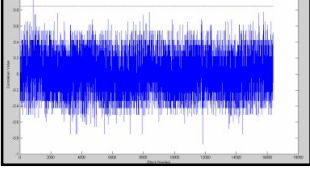
| | | |
|--|---|---|
|  MEDIAN_3 |  |  Max corr value 0.88 |
|  RML_90 |  |  Max corr value =1 |
|  ROTCROP_-75 |  |  Max corr value 1 |
|  ROTSCALE_0.5 |  |  Max corr value 1 |

TABLE 3. Performance Evaluation

5. REFERENCES

1. Er-Hsinen, "Literature Survey on Digital Image Watermarking," *EE381K-Multidimensional Signal Processing* 8/19/98.
2. S. Katzenbeisser and F. A. P. Petitcolas: *Information Hiding techniques for steganography and digital watermarking*, Artech House, Inc., MA, USA, 2000.
3. N. Memon and P. W. Wong.:Protecting Digital Media Content. *Communications of the ACM*, vol. 41, no. 7, pp. 34–43, Jul 1998.
4. C.C. Chang and J. C. Chuan, "An image intellectual property protection scheme for gray-level images using visual secret sharing strategy," *Pattern Recognition Letters*, vol. 23, pp. 931-941, June 2002.
5. S. P. Mohanty.:Watermarking of Digital Images.M.S. Thesis, Indian Institute of Science, Bangalore, India, 1999.
6. N. Nikolaidis, I. Pitas, "Robust Image Watermarking in Spatial Domain", *International journal of signal processing*, 66(3),385-403,1998.

7. Pankaj U. Lande, Sanjay N. Talbar and G.N. shinde "Adaptive DCT Domain Watermarking For Still Images", *Internatational Conference RACE-07*, Bikaner, Rajasthan, India
8. Pankaj U. Lande, Sanjay N. Talbar and G.N. shinde "Hiding A Digital Watermark Using Spread Spectrum At Multi-Resolution Representation", *Internatational conference ACVIT07*, Aurangabad, India.
9. Juan R. Hernandez, Martin Amado, Fernando Perez-Gonzalez "DCT Domain watermarking technique for still Image :Detectors Performance analysis and New Structure", *IEEE transaction on image processing*, VOL.9, no.1, Jan 2000.
10. Juan R. Hernandez, Martin Amado, Fernando Perez-Gonzalez "DCT Domain watermarking technique for still Image: Detectors Performance analysis and New Structure", *IEEE Transaction on Image Processing*, VOL.9, No.1, Jan 2000.
11. Pik Wah Chan, Michael R. Iyu and Roland T. Chin, "A Novel scheme For Hybrid Digital Video Watermarking : Approach, Evaluation And Experimentation", *IEEE Transactions on circuits and system for video technology*, VOL 15, No. 12, Dec 2005.
12. Pankaj U. Lande, S.N. Talbar, G.N. Shinde, "FPGA implementation of image adaptive watermarking using human visual model", *ICGST-PDCS*, Vol.9, Issue1, Oct. 2009.
13. I. J. Cox, J. Kilian, T. Shamoan, T. Leighton, Secure Spread Spectrum Watermarking of Images, Audio and Video, in: Proc IEEE International Conf on Image Processing, Vol. 3, 1996, pp. 243–246.
14. I. J. Cox, J. Kilian, T. Shamoan, T. Leighton, A Secure Robust Watermarking for Multimedia, in: Proc. of First International Workshop on Information Hiding, Vol. 1174, 1996, pp. 185–206.
15. P. Meerwald and A. Uhl, (2001) "A survey of wavelet-domain watermarking algorithms," Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents III, San Jose, California, vol. 4314.
16. Burrus, C., Gopinath, R., and Guo, H.: *Introduction to Wavelets and Wavelet Transforms: A Primer*.: Prentice Hall 1998.
17. R. Dugad, K. Ratakonda, and N. Ahuja, (1998) "A new wavelet-based scheme for watermarking images," *Proceedings of the IEEE International Conference on Image Processing, ICIP '98*, Chicago, Illinois, pp. 419-423.
18. D. Taubman and M. Marcellin.: *JPEG2000: Image compression fundamentals, standards, and practice*.: springer, 2002.
19. T. Ebrahimi and F. Pereira.: *The MPEG-4 Book*.: Prentice Hall 2002.
20. Y. H. Seo, D. W. Kim, Real-Time Blind Watermarking Algorithm and its Hardware Implementation for Motion JPEG2000 Image Codec, in: Proceedings of the 1st Workshop on Embedded Systems for Real-Time Multimedia, 2003, pp. 88–93.
21. S. P. Mohanty, R. K. C., S. Nayak, FPGA Based Implementation of an Invisible-Robust Image Watermarking Encoder, in: Lecture Notes in Computer Science, Vol. 3356, 2004, pp. 344–353.

22. S. P. Mohanty, O. B. Adamo, and E. Kougianos, "VLSI Architecture of an Invisible Watermarking Unit for a Biometric-Based Security System in a Digital Camera", in *Proceedings of the 25th IEEE International Conference on Consumer Electronics (ICCE)*, pp. 485-486, 2007.
23. S. P. Mohanty, "A Secure Digital Camera Architecture for Integrated Real-Time Digital Rights Management", *Elsevier Journal of Systems Architecture (JSA)*, Volume 55, Issues 10-12, October-December 2009, pp. 468-480. Kanchan H. Wagh, Pravin K. Dakhole, Vinod G. Adhau.: Design & Implementation of JPEG2000 Encoder using VHDL. Proceedings of the World Congress on Engineering 2008 Vol I, WCE 2008, , London, U.K July 2 - 4, 2008.
24. Kanchan H. Wagh, Pravin K. Dakhole, Vinod G. Adhau.: Design & Implementation of JPEG2000 Encoder using VHDL. Proceedings of the World Congress on Engineering 2008 Vol I, WCE 2008, , London, U.K July 2 - 4, 2008.
25. M. Kutter and F.A. Petitcolas, "A Fair Benchmark for Image Watermarking Systems", *Electronic imaging , Security and Watermarking of Multimedia Contents, VOL. 3657*, 25-32, 1999.

Comparing the Proof by Knowledge Authentication Techniques

Stamati Gkarafli

*Information Systems Department
University of Macedonia
Thessaloniki 54006, GREECE*

gkaraflistamati@yahoo.gr

Anastasios A. Economides

*Information Systems Department
University of Macedonia
Thessaloniki 54006, GREECE*

economid@uom.gr

Abstract

This paper presents a survey of proof by knowledge authentication techniques (text passwords, visual passwords and graphical passwords). Both new methods are more memorable, as people have to remember images and not characters and graphical passwords are also more secure. A total of 100 users participated in our survey, who after getting informed about the new authentication methods, they answered the questions of our questionnaire. Based on their answers, all participants have many passwords for their everyday needs and they try to select passwords that are not only memorable, but also secure. Unfortunately, they can not deal with proper password selection and they become victims of dictionary attacks. Understanding this situation, participants were very positive in learning more about the new authentication methods. They found both techniques memorable and friendly – visual passwords at most. However, they found graphical passwords a bit more complex and difficult to learn how to use them, something that they can overcome with small practice.

Keywords: graphical passwords; text passwords; user authentication methods; visual passwords.

1. INTRODUCTION

Access control includes the user's identification and authentication, authorization, audit and accountability. Various access control models have been proposed in the past [e.g. 1-5].

Authentication is the process of confirming or not the user's identity. Jansen [6] distinguished the authentication techniques into the following three categories:

Proof by knowledge techniques, which are based on specific information that an individual has (e.g. PIN- personal identification number, text-passwords).

Proof by property techniques, which are based on a certain property that the user has (e.g. biometrics, fingerprint verification, voice verification, iris scanning) [e.g. 7].

Proof by possession techniques, which are based on the possession of an object that an individual has (e.g. smart cards, digital certificates, security token).

Text passwords represent the authentication method that is mainly used by all users today. Nevertheless, most times users select passwords that are memorable and as a result easy to be cracked [8]. This problem is very serious if one looks at some case studies that were conducted. According to a security team in a large company, they managed to crack 80% of the passwords [9]. Also, based on Klein's case study [10], 25% of 14.000 passwords were cracked using a small dictionary of 3 million words. According to these results one can assume that even if these methods

are very popular, they can cause serious problems to the users. Some of the usual problems are the following:

- Users choose passwords that are very short in length.
- Users choose passwords that are easy to remember.
- Users write passwords down or share them with others, in order to remember them easier.
- Users use the same passwords for different applications.

Text passwords are very vulnerable to “dictionary attacks” (automated attacks using tools that can crack the passwords that are common words, names or dates).

2. VISUAL AND GRAPHICAL PASSWORDS

Considering all those problems of text passwords, researchers invented other proof by knowledge authentication methods: visual and graphical passwords. These methods have many advantages that are described below [11, 12, 13]:

A sequence of pictures is more memorable than a sequence of characters.

Pictures are independent from user’s language.

There do not exist yet special dictionaries for a dictionary attack and it is very difficult to be constructed (especially for graphical passwords that have a very large password space).

Automated attacks are difficult to take place.

Except from all these advantages Renaud and De Angeli [14] referred the shoulder surfing problem as the main disadvantage of the new methods. This happens when someone is looking over someone’s shoulders, during the login process [15, 16].

Next, we describe the two new authentication methods and the applications that have been created for each one of them.

2.1 Visual Passwords

Visual passwords are passwords that are created by selecting a sequence of images [5, 17]. Based on this idea, there was developed a number of commercial products that we will refer and analyze in the text below.

Passfaces

Real User Corporation developed a product named “Passfaces” [14, 18]. The basic idea is that the user must select four images of human faces, in a specific order, from a database. This sequence of images will be his secret password.

During the login process, the user sees a grid of nine faces, but only one of those faces belongs to the password that he made. He clicks on it, another set of nine images appears and he must again recognize and click on the image that is included in his password. This procedure is repeated until the user clicks on all four pictures of his password. If he clicks on all four pictures correctly, the system allows him to enter.

The whole idea is based on the fact that people recall much easier human pictures than any other kind. This is especially true if each user has the opportunity to construct his own database with his personal pictures which are familiar to him and much harder to be guessed by an attacker.



FIGURE 1: Passfaces

Story Scheme

Story Scheme is an application similar to Passfaces. The only difference is that the images would be not only human faces, but also everyday objects, animals, food, sports, cars, or even people. In this case, the user has to think a story with the pictures that are in his database. Having his story in his mind, he selects a sequence of images [19].

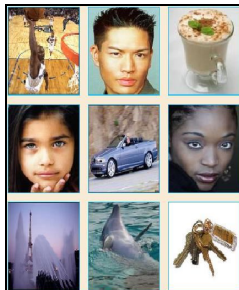


FIGURE 2: Story Scheme

Déjà Vu

Dhamija and Perrig developed “Déjà Vu” for user authentication [20]. At the beginning the user creates his “image portfolio”, by selecting a set of p images out of a much bigger set.

During the login process, the Déjà Vu presents to the user a set of n images consisting of: x images that belong to the users’ image portfolio, and y other images, that we call “decoy images”.

At this point the user has to make a click on all images that he recognizes as belonging into his image portfolio. If he succeeds in it, the user will enter the system successfully.

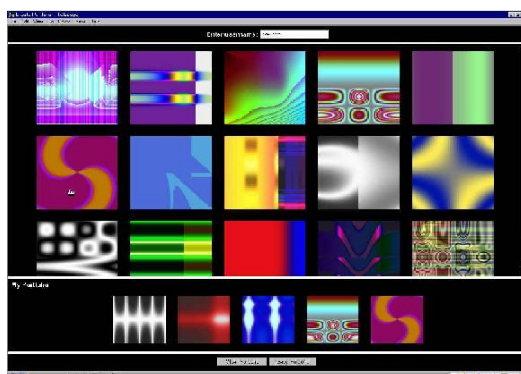


FIGURE 3: Déjà Vu

Something that is very important in Déjà Vu scheme is the type of images that are used. The images are based on Andrej Bauer’s Random Art [21], which can generate random abstract images. When the user attempts to make his image portfolio, the system gives to this specific process an initial seed. With this seed, Random Art generates a random mathematical formula, which defines the color of each pixel in an image. Therefore, each user has his personal random pictures that were created from the initial seed.

So, it is obvious that this type of images makes the whole system much more secure, since it is very difficult for someone who observes the login process to remember them. Moreover, the system does not have to store each image pixel-by-pixel, but it stores only the initial seed.

Picture Password

Picture Password authentication mechanism is another application for visual login, which was developed by Jansen [5, 22].

The images are grouped into different categories according to the theme that they represent. Theme examples include Cats & Dogs, Sea, Landscapes, Sports, Faces, Transportation Means, etc. In order to create a password, the user has to choose one of these themes and afterwards a sequence of images from this theme.

Each image corresponds to an element of an alphabet. However, the user does not have to remember a sequence of random characters, but a sequence of images, something that is much easier to recall.

There are two different ways to choose the sequence of images:

Individual Selection: each image represents one element in the alphabet

Paired Selection: two images are combined (more often with drag and drop) and their coupling represents one element in the alphabet.



FIGURE 4: Picture Password.

Passlogix – Passpoints

Passlogix [23] is based on Blonder's idea [24] who proposed that during authentication, the user must click on several locations in an image. In the Passlogix implementation, the user must click on a sequence of items in the image he sees on his screen in order to create his password. To make a successful login to the system, he has to click on the same items in the correct order. A problem that we face here is how we know if an item is clicked, if we click on its edge. For this reason, there were defined invisible boundaries which indicate whether an item is clicked by the mouse or not.

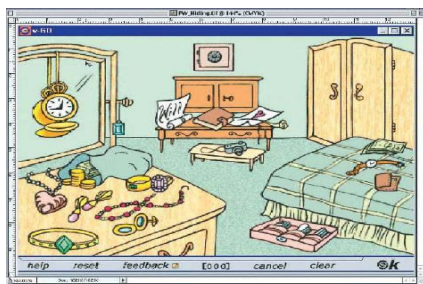


FIGURE 5: Passlogix.

An extension of this idea is the "Passpoint" system developed by Wiedenbeck et al. [25, 26, 27]. In this implementation, the user is able to make a click at any place on an image, and not only on a certain object. To achieve this they eliminated the boundaries between objects and put a tolerance around each pixel that is selected. As a result, the only thing that a user must do to make a login and enter the system is to click within the tolerance of the pixels that he chooses, in the correct order.

It is important to be mentioned here that complex pictures can have hundreds of memorable points. So, with 5 or 6 clicks, a user can make more passwords than that used today with 8 characters. So, the possible password space with this method is very large, and the whole process becomes very safe.

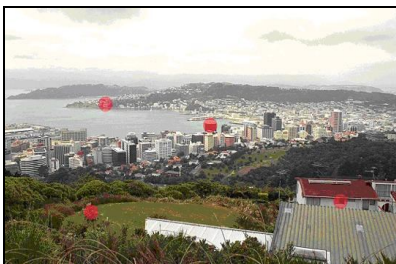


FIGURE 6: Passpoints.

2.2 Solving the Shoulder Surfing problem

As we have already mentioned, one main problem that visual passwords have is the “shoulder surfing problem”. That is watching over someone’s shoulders as he tries to login a system [16]. Next, we present three methods developed by Sobrado and Birget [15] in order to solve this problem.

Triangle Scheme

According to this scheme, the user sees on his screen a set of N objects in a random position each time. He creates his password by selecting K pass-objects that will consist from now on, the user’s portfolio.

During the login process the user is able to see on his screen a set of L images with $L < N$. To login correctly he must recognize the 3 pass-objects that are depicted and make a click inside the invisible triangle that is created.

If this process is done only once, then it would be very easy for someone to click inside the triangle by chance. So, the process is repeated for 8 – 10 times for each login, with the purpose to reduce the probability of randomly clicking on the correct region.

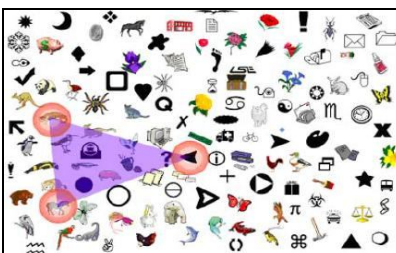


FIGURE 7: Triangle Scheme.

Movable Frame Scheme

Based on the same ideas with the previous scheme, in the “Movable Frame Scheme”, the user has to recognize 3 of the pass objects that he has already chose. This time, his job is to move the frame around (Figure 8) until the pass object on the frame lines up with the other two.

Of course, the process is again repeated for several times, as it was also mentioned above, in order to avoid lining up the correct items by chance.

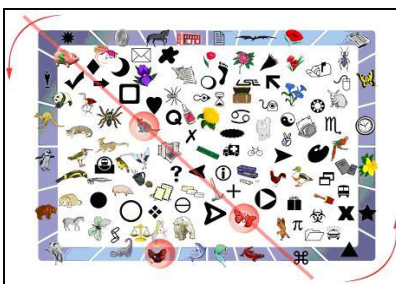


FIGURE 8: Movable Frame Scheme.

Other Geometric Configurations

Similarly, the third method follows the same rules. However, in this case the user has to recognize 4 pass objects and make a click at the intersection of the virtual lines that are formed by connecting these objects.

Of course, it is obvious, that there is a tolerance of the pixel that the user clicks, in order not to have arguments about the exact point that he chooses each time.

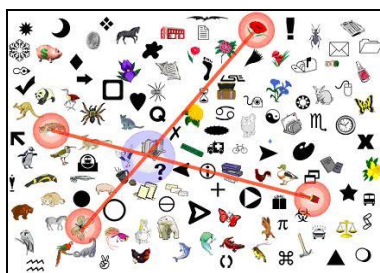


FIGURE 9: Other Geometric configurations.

2.3 Graphical Passwords

In this section, we analyze graphical passwords. Here, the user has to draw a personal design that will be from now on his secret password. More precisely, we examine the DAS scheme and its extension Multi-grid Passwords.

Draw-a-Secret (DAS) Scheme

Jermyn et al. [22] proposed a scheme, called DAS (Draw-a-Secret), which allows the user to draw a simple design on a grid. On his screen the user is able to see a rectangular grid of size $G \times G$. Then, he has to draw a sequence of lines in this grid. This drawing will represent his password.

For example, let consider a grid of size 3×3 (Figure 10). In order to create his password, the user has to draw a sequence of lines. Then, at the login process, he must draw the same lines again, in the same order. For this reason the drawing is mapped to a sequence of coordinate pairs by making a list of the cells through which the drawing passes in the correct order, separated by pen-up events, when the user raises his pen and continues from another point. In Figure 10, we have the following sequence that we made by the drawing:

(1,2), (1,1), (2,1), (2,2), (3,2), pen-up, (2,3), (1,3).

According to this, we must underline that what counts for a user in order to make a successful login is not the exact draw that he has made, but the sequence of cells from which his pen passes in combination with the pen-up events.

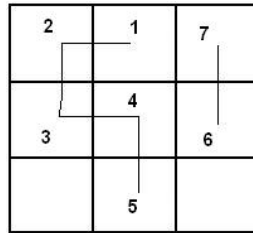


FIGURE 10: DAS Scheme.

In Figure 11 we can see another example of the DAS Scheme, its internal representation, the bit string that is created and an unsuccessful login because of a shift error that was made.

An important point is that a one-way hash function is applied to the bit string that is created and the result is stored to the server of the system. So, when the user tries to login the system, he draws his password, then the hash function is applied to it, and the result is compared with the stored result. If these two are the same, then the login is successful, otherwise it is not.

As we can understand this process makes the whole scheme very safe. The system does not know each user's password and as the hash function is one-way we cannot compute the initial design from the result that is stored.

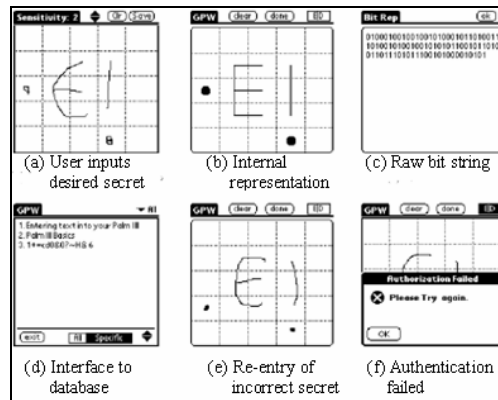


FIGURE 11: DAS Scheme. Internal representation of our draw and a failed login because of a shift error [28].

a) Multi-grid Passwords

Nali and Thorpe [29] performed a survey to investigate the reasons that could make DAS passwords vulnerable to attacks. According to them, this idea could be true if we consider that many users draw passwords with predictable characteristics, which means that these are centered or symmetric. As a result symmetrical or centered DAS passwords can reduce very much the password space and help attackers in creating dictionaries [30, 31]. Based on this remark, Birget et al. [30] referred problems with the DAS scheme because of uncertainty in the clicking regions.

Chalkias et al. [33] as well as Alexiadis et al. [34] made an extension to the DAS scheme with the aim to reduce these facts that have very serious implications to the security of the system. They proposed multi-grid passwords by using nested grids in the initial one (Figure 12). The main reasons that users fail to confirm their passwords are that they forget their stroke order and they mark adjacent cells instead of the correct ones.

With multi-grid passwords the users are able to create more complicated passwords that at the same time are more memorable. For example, using grids like in Figure 12, the user has more points of focus to do his drawing, and so he does not need to make his design in the center of the grid or symmetrically. With multi-grid passwords the users are able to create more complicated passwords that at the same time are more memorable. This can be explained if we examine Figure 12. With grids

like those, we have more points of focus to do our drawing, and so the user does not need to make his design in the center of the grid or symmetrically.

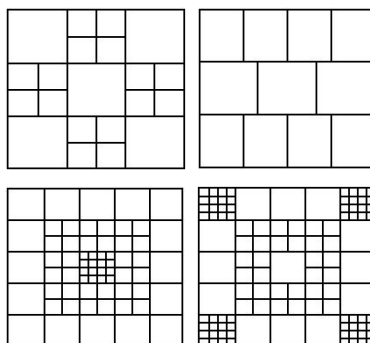


FIGURE 12: Multi-grid passwords [33].

3. RELATED WORK

As we have already mentioned, text passwords nowadays are very unsafe and can cause serious security problems to the users. Based on this fact and taking into account the new authentication methods that were invented, many researchers have worked on them, with the aim to compare them and find what effect they may have in our society.

According to Irakleous et al. [35], 59% of the participants have 2 -5 passwords, while 65% of them use at least one of them daily. Also, Tribelhorn [36] showed that all users are concerning very much for the security of their passwords. For this reason, many users try to create passwords that are difficult to be cracked, by selecting 8 or more characters. But just this effort is not enough, because users want their passwords to be also memorable. So they create passwords that are consisted of ordinary words or dates, making them really vulnerable to dictionary attacks.

All these lead the researchers to the new authentication methods, visual and graphical passwords. Many applications were created based first of all on the image selection of visual passwords. Kim and Kwon [37] found that having in visual passwords images with known faces, make them more memorable that having landscapes or random faces. Davis et al. [19] examined the images that males and females choose and concluded that females chose animals or food and males choose women and sports. Dhamija and Perrig [20] working on Déjà vu scheme, found out that after some practice, users have very good results in remembering the passwords with simple photos, but also the passwords that are based on random art technique.

Really opposed to these results, Jansen [22] showed that visual passwords are not safer than the text, as users tend to select a small number of images (usually 4 or 5), creating this way passwords easy to be cracked. Wiedenbeck et al. [25, 26, 27] showed that graphical passwords may be more time-consuming or more difficult to be confirmed successfully, but that after some practice the situation changes and above all the passwords that are created are really safe.

Further examining the graphical passwords, Goldberg et al. [38] found out that with practice users can have the best results. They make more successful logins, something that is confirmed by the fact that 72% of them agreed that this method is easier to remember than the text passwords.

Nali and Thrope [29] examined the drawbacks of graphical passwords which include that users draw designs that are centred or symmetrical, something that decreases their security. As a solution to this situation, Chalkias et al. [33] proposed a multi-grid password scheme and they made a comparison for centred and symmetrical drawings, between non-technical and technical users. According to their survey, using a multi-grid scheme fewer participants created centred and symmetrical passwords and as a result, more users were able to create safer passwords. Weiss and Del Luca [39] proposed PassShapes. In this system users authenticate themselves to a computing system by drawing simple geometric shapes constructed of an arbitrary combination of eight different strokes. Also, Eljetlawi and Ithnin [40] designed Jetafida focusing on the usability features of this graphical password system. Everitt et al. [41] found that the frequency of access to a graphical password, interference resulting

from interleaving access to multiple graphical passwords, and patterns of access while training multiple graphical passwords significantly impact the ease of authenticating using multiple facial graphical passwords. Chiasson et al. [42] compared the recall of multiple text passwords with recall of multiple click-based graphical passwords. In a one-hour session (short-term), they found that participants in the graphical password condition coped significantly better than those in the text password condition. Similarly, Ozok and Holden [43] compared alphanumeric and graphical passwords. Johnson and Werner [44] found that passcodes are more memorable than alphanumeric passwords over extended retention intervals. Finally, various classification systems of graphical passwords were proposed [45, 46].

In our survey, in opposition to those that we have already mentioned, we made a comparison between the three “proof by knowledge” authentication techniques, text, visual and graphical passwords and not just between two or three specific applications that are listed either in the visual or graphical method. Also 100 users were participated with the aim to obtain results that are significant statistically, while in the other surveys there were participated up to 40 users. Finally, to have more general results, the users that participated were from 18 to 60 years old (not only students in colleges and universities), in order to find out the impression and the effect that the new methods have in the whole society and not just in a part of it.

Furthermore, we created a comprehensive study, to examine many issues that refer to visual and graphical passwords, comparing them with the traditional authentication method that is text passwords.

4. METHODOLOGY

4.1 Questionnaire

To conduct our survey, we created a questionnaire to obtain users' opinion. Our questionnaire is a completely new questionnaire that is referred to characteristics that text, visual and graphical passwords have. More precisely, the information that we collected with the questionnaire are the following:

how important are passwords for the users

what problems do text passwords have

are they positive in using visual and graphical passwords

user's personality according to their personal opinion

users' opinion about text, visual and graphical passwords in terms of memorization, difficulty in learning their use and friendliness.

Analyzing users' answers to the questionnaire, we discovered their opinion about the traditional text passwords and the new visual and graphical passwords. In addition, we examined at what degree our society is able to change its habits and accept without any problems, the new authentication methods that are really safer.

4.2 Participants

In our study 100 people were participated. They already knew and used PINs and text passwords. There were included 49 men and 51 women, from 18 to 60 years old. More precisely, if we divide the users into three categories according to their age, we have: 59 users from 18 to 30 years old, 20 users up to 45 years old and 21 users over 46 years old and until 60. The younger participants were students or post graduated students in the University of Macedonia in Thessaloniki, in the Aristotle University in Thessaloniki and in the TEI of Larissa. The elder ones were teachers in schools and Universities, people that work in public services, in the private sector, in manufactures or in factories, in Thessaloniki and in Larissa. All these users were found in their work, something that made really difficult the whole process of collecting their answers. This is because we had to make the procedure that we will describe analytically below, separately to 3 - 4 users, as it was infeasible to find more people free at the same time, to deal with our questionnaire.

What was very important to us was that from the beginning of our survey every participant was very positive in helping us to carry it out. Especially after listening to a small introduction with information about the new methods and the danger that traditional text passwords hide, they were very willing to participate and in this way help us overcome all these problems.

4.3 Procedures

As we have already mentioned, in order to accomplish our survey, we have created a questionnaire that referred to text, visual and graphical passwords. The whole procedure of our survey is divided into two different stages. At the first stage, we gave a short seminar (as many of them were in their work) to the participants that included an introduction to the text passwords and their problems and an analysis of the new authentication methods, visual and graphical passwords.

After understanding all these points, all the participants were asked to fill in the questionnaire that we have prepared, with the aim to find out their personal opinion about the three authentication methods. Finally, after collecting all the questionnaires, we analyzed the users' answers and we present the results in the text below.

5. RESULTS

In this section, we present the results of the survey regarding text, visual and graphical passwords. We present the results in six categories: 1) password importance, 2) problems with text passwords, 3) the new authentication techniques, 4) comparison among the three methods, 5) making safer graphical passwords.

5.1 Are passwords important to our life?

First of all, we were interested in identifying the importance of passwords from the users' point of view. We asked participants about the number of passwords they use in their everyday activities. 47% of them have one to four passwords, 34% have 5 to 7 passwords, and 19% have more than 8 passwords.

In Figure 13, we present the percentage of men and women with respect to their number of passwords. Almost the same percentage of men and women has one to four passwords. However, 26.53% of men and 41.18% of women have five to seven passwords. On the contrary, 26.53% of men and 11.78% of women have eight or more passwords. So, we observe a significant difference between men and women who use more than four passwords.

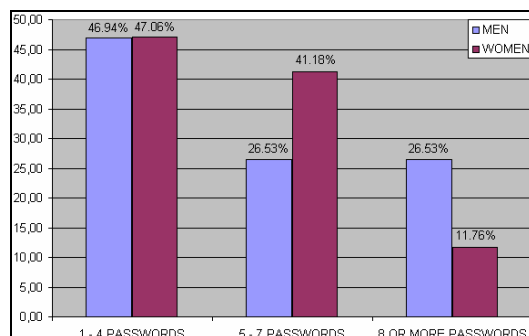


FIGURE 13: Number of passwords that men and women use

Taking into consideration the age of the users, we observe that the majority (76.19%) of people between 46 and 60 years old use up to four passwords, while younger people tend to use many passwords (Figure 14). Five to seven passwords are used by 35.59% of people between 18 and 30 years old, and by 40% of people between 31 and 45 years old. More than eight passwords are used by 27.12% of people between 18 and 30 years old, and by 15% of people between 31 and 45 years old.

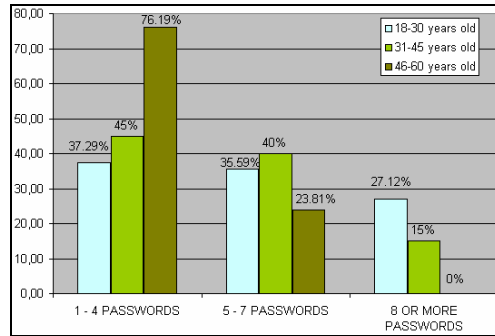


FIGURE 14: Number of passwords that users have, according to their age.

It is also important to examine for what reasons they use those passwords, and how often they really use them. Almost every user has a mobile phone (94%) and ATM cards (88%) with which he uses passwords to verify his identity (Figure 15). Considering the increased use of Internet, we can understand why so many people have passwords for their e-mail (64%), Internet connection (54%) or to make a login to web pages (39%).

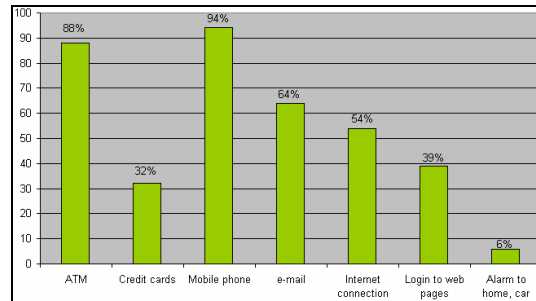


FIGURE 15: The applications where users use passwords.

Finally, we found that 69% of the participants use at least one password every day, and almost all of them use a password once or twice per month. So, we can figure out how important is for everyone to have a memorable and primarily safe password.

5.2 Analyzing text passwords

As we had already referred, text passwords nowadays are very unsafe and can cause many problems to the people that use them. To find out if this situation is really true, we asked the participants what kind of passwords they use for their applications. Their answers are presented in Figure 16.

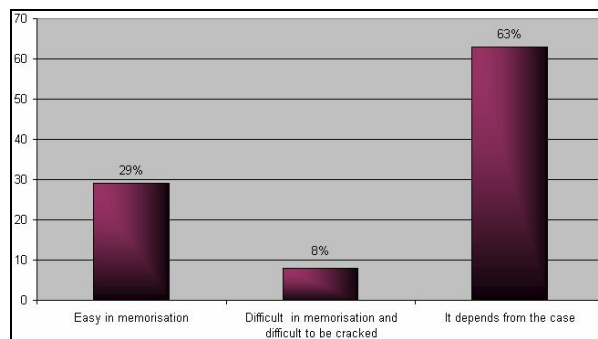


FIGURE 16: Type of passwords that users choose.

The percentage of the users that choose passwords that are easy in memorization is 29%, while just 8% of them choose passwords that are difficult in memorization but also difficult to be cracked and as a result very safe. The rest 63% of the participants chooses both easy and difficult passwords, based on how important are the applications that are applied to.

These percentages are quite good at a first glance. But to make things more clear and to understand if these results are really satisfactory, we have to find out more details about these passwords.

In Figure 17, we can see the number of characters that participants use in their passwords. 41% of them have 4 characters in their passwords, 43% have 5 to 7 characters and 16% have more than 8 characters.

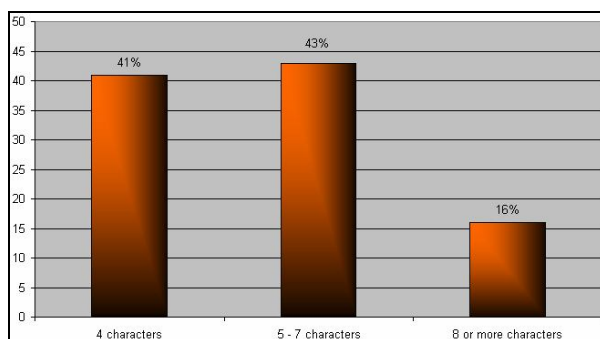


FIGURE 17: Number of characters in a text password.

Carefully looking at these results we can say that they are quite satisfactory as a password of more than 8 characters is characterized very safe and really difficult to be cracked and one with 6 or 7 characters is characterized as a password of medium difficulty. But these results are incomplete, because we have not examined yet what is the exact kind of the characters that these passwords are consisted of. For this reason we have created Figure 18.

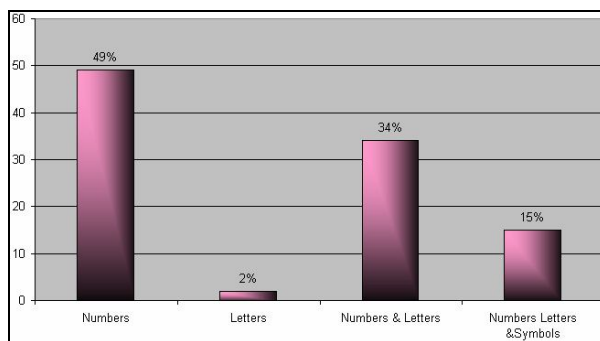


FIGURE 18: The exact kind of characters that the passwords are consisted of.

According to Figure 18, we can see that 49% of the participants use just numbers in their passwords and 2% just letters. This situation is really problematic because most participants that use just numbers or letters in their passwords create passwords with familiar dates or names that consequently are very vulnerable to attacks. Only 15% of these passwords is consisted of numbers, letters and symbols together and are really safe and reliable passwords.

5.3 New authentication techniques: Visual and Graphical passwords

Considering the previous results, we can be sure that most text passwords that were created by users are predictable and really unsafe for them. So, it was inescapable that new authentication methods were needed. These methods are visual and graphical passwords.

Next, after explaining to the users about visual and graphical passwords, we examine at what degree they are positive in further learning and using them. Their answers are divided into five categories (Figure 19), that are “by no means”, “a little”, “enough”, “much”, “very much”.

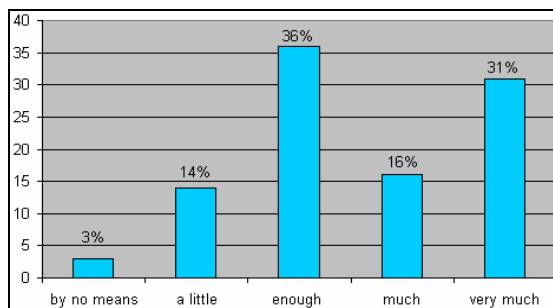


FIGURE 19: Users’ position in accepting or not of visual and graphical passwords

As we can observe in Figure 19, the users that are totally negative or a little interested in learning more about the new authentication methods are very few (3% and 14% correspondingly). In opposition to this, 36% of the users chose “enough” as their answer, 16% “much” and 31% “very much”, a situation that reveals that most of the participants in our survey are positive in learning and also using visual and graphical passwords.

5.4 Analyzing the three authentication methods

In this section we will analyze the two new authentication methods, visual and graphical passwords, with respect to various parameters that will show us if these methods affected participants in a positive way.

Firstly, we will examine the memorability of each method. As we can see in Figure 20, 57% of the users believe that visual passwords are more memorable, while 43% of them prefer graphical passwords.

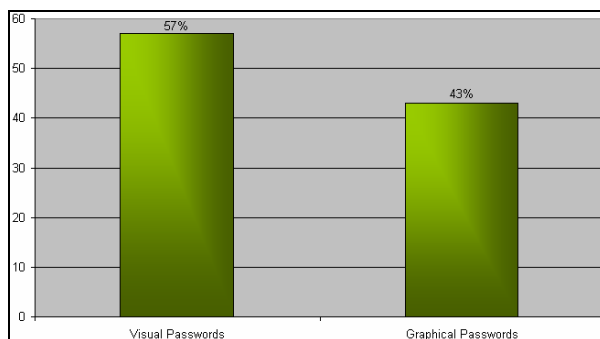


FIGURE 20: Most memorable authentication method

Based on these results we found it very interesting to divide the participants into three categories, describing their personality according to their personal opinion. To answer in this question the participants had to select one out of the three choices that follow:

Visual, if the user remembers or learns something easier using images

Acoustic, if the user remembers or learns something easier by listening to it

Verbal, if the user remembers or learns something easier, when it is written down.

According to their answers, we conclude in Figure 21.

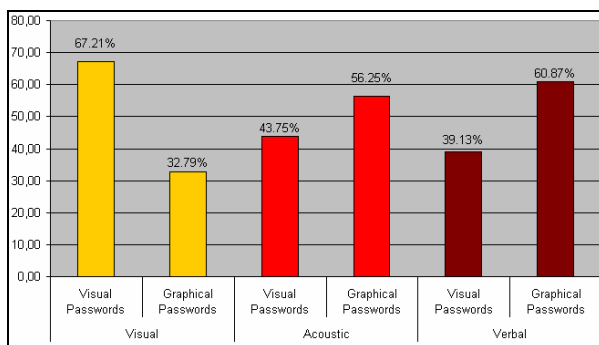


FIGURE 21: Most memorable authentication method based on the kind of person that each user enlists himself.

Here we can see that more users that are visual types of persons prefer Visual Passwords (67.21%), while users that are acoustic and verbal types of persons, prefer mostly Graphical Passwords (56.25% and 60.87% correspondingly).

Next, we will examine and compare all three authentication methods (text, visual and graphical passwords) regarding users’ opinion about how easy is it for them to learn how to use each method. Users had to select one out of the five answers that were: “not at all”, “a little”, “enough”, “much” and “very much”, in proportion to how difficult is the whole process for them. Collecting their answers, we conclude in Figure 22.

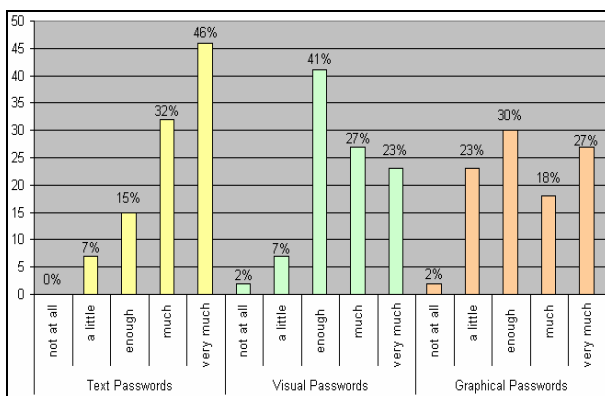


FIGURE 22: How easy is for someone to learn how to use the three authentication methods

According to Figure 22, we can observe that the users that found it very difficult or a little difficult to learn how to use text and visual passwords are very few. Most of them found the two methods really easy, for two different reasons:

- Text passwords, because they are very familiar with them, as they use them almost every day
- Visual passwords, because they are attractive and they have no special rules (the users just have to choose a sequence of images as their password).

On the other side, only 2% of the users find it very easy to learn how graphical passwords are used and 23% of them believe that the whole process is a little easy to be learnt. This percentage is not really problematic, as this method is not so attractive at first side and has also many rules that a user must remember, to create his password. What we have to mention here is that the rest of the participants did not have difficulties in learning the use of graphical passwords and above all they believe that with some practice everyone will be able to overcome any problem he may face.

The last parameter that we examined is how easy is for someone to use the three authentication methods. Users here should again select one of the five given answers that are “not at all”, “a little”,

“enough”, “much” and “very much”, regarding how friendly they think that each method is for them. The results are depicted in Figure 23.

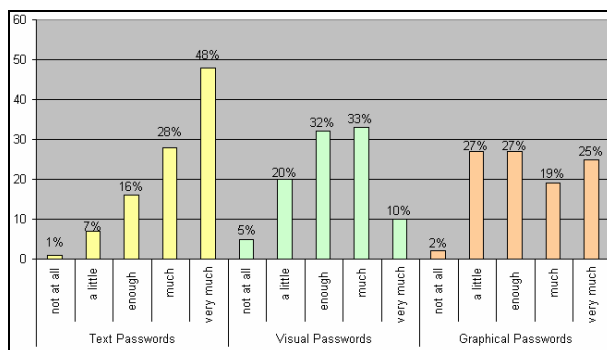


FIGURE 23: User friendliness of the three methods

In Figure 23, we can see that text passwords are considered to be the friendliest method. For visual and graphical passwords 5% and 2% correspondingly believe that they are not friendly at all to the users, 20% and 27% that they are a little friendly, while the rest believe just the opposite. Moreover, we have to say that 25% of the users chose “very much” as their answers for graphical passwords, while the same percentage for visual is 10%. This situation proves us that even if visual passwords impressed the users more at the beginning of the survey, the participants were able to overcome all the difficulties and understand why graphical passwords is the best method out of the three.

5.5 Making more safe graphical passwords

After finishing processing the data, analyzing the results and taking into account other researchers’ results, we conclude that graphical passwords is the safest authentication method. Moreover with some practice by the users, they can become very memorable and ease to be used by the users.

Considering all these, we suggest a set of tips and advices to users, in order to create very difficult passwords and not vulnerable to attacks. These tips are the following:

- do not make symmetrical shapes
- do not make centered shapes
- try to draw more than one lines
- include in your drawing, as much pen up events as you can (the same drawing with more pen up events is much safest)
- avoid starting your drawing from the 4 corners (these cells are very vulnerable)
- if your drawing is simple, try to make a second or even a third, similar to the initial one but covering different cells
- avoid drawing diagonal lines, and as a result lines near the intersection of the lines that create the cells, because it is very easy to get confused and make shift errors.

6. CONCLUSIONS

We conducted our survey with the purpose to compare the proof by knowledge authentication techniques which are text passwords, visual passwords and graphical passwords. A total of 100 users participated in the survey. First, they were informed about the problems that text passwords have and the advantages that visual and graphical passwords can offer to overcome these problems. Finally, according to what they learnt and their personal opinion, they answered to our questionnaire.

Based on users’ answers we are able to confirm that almost every user uses passwords for different applications such as mobile phone, ATM cards, credit cards, e-mails, internet connection etc. From these users, almost half of men and women have 1 – 4 passwords, while about 40% of women have 5 – 7 and 26% of men 8 passwords or more. Indeed, as it was expected, users that have 1 – 4 passwords are the elder ones (46 – 60 years old) and those that have more passwords are the younger.

Analyzing the text passwords that users claim that they create, we see that 30% of them create passwords that are really memorable and as a result easy to be cracked. The encouraging thing is that 63% of them create either easy or difficult passwords (with as many characters as they can), with respect to the application that they are referred to. Unfortunately, this is not enough because many users include only numbers in their passwords, or only letters, or numbers and letters both, creating passwords of familiar dates and names that are vulnerable to dictionary attacks.

That was the main reason that a great percentage of users were very positive in knowing better the new authentication methods. After learning more information about these methods, 57% of them chose visual passwords as the most memorable method, while 43% preferred the graphical passwords. Besides, as we divided the users in three categories, visual, acoustic and verbal, we must say that visual users prefer mainly visual passwords while acoustic and verbal users prefer graphical passwords.

Moreover, we must report that both new methods and especially visual passwords were characterized very friendly by all users. At the same time graphical passwords were characterized a bit difficult until they learn how exactly they are used.

Finally, keeping in mind that graphical passwords is the safest method, we made a small list with tips that users must follow, in order to create really difficult graphical passwords and as a result very hard to be cracked from anyone.

REFERENCES

- [1] Bammigatti, P. H., and Rao, P. R. "Delegation in role based access control model for workflow systems". *International Journal of Computer Science and Security*, 2(2): 1-10, 2008.
- [2] Chandrasekar, A., Rajasekar, V. R. and Vasudevan, V. "Improved authentication and key agreement protocol using elliptic curve cryptography". *International Journal of Computer Science and Security*, 3(4): 325-333, 2009.
- [3] Kar, J. and Banshidhar, M. "An efficient password security of multi-party key exchange protocol based on ECDLP". *International Journal of Computer Science and Security*, 3(5): 405-413, 2009.
- [4] Tahir, M. N. "Hierarchies in contextual role-based access control model (C-RBAC)". *International Journal of Computer Science and Security*, 2(4): 28-42, 2008.
- [5] Tahir, M. N. "Testing of contextual role-based access control model (C-RBAC)". *International Journal of Computer Science and Security*, 3(1): 62-75, 2009.
- [6] W. Jansen. "Authenticating users on handheld devices". In *Proceedings of the Canadian Information Technology Security Symposium*, 2003.
- [7] Bhagwat, R. and Kulkarni, A. (2010). "An overview of registration based and registration free methods for cancelable fingerprint template". *International Journal of Computer Science and Security*, 4(1): 23-30, 2010.
- [8] H. Davies. "Physiognomic access control". *Information Security Monitor*, 10(3): 5-8, 2005.
- [9] K. Gilhooly. "Biometrics: Getting back to business". *Computerworld*, May 2005.
- [10] D. Klein. "Foiling the Cracker: a survey of, and improvements to, password security". In *Proceedings of the 2n USENIX Security Workshop*, pp. 5-14, 1990.

[11] de A. Angeli, L. Coventry, G. Johnson and K. Renaud. "Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems". *International Journal of Human-Computer Studies*, 63: 128-152, July 2005.

[12] H. Bolande. "Forget passwords, what about pictures?". <http://zdnet.com.com/2102-11-525841.html>

[13] X. Suo, Y. Zhu and S. G. Owen. "Graphical passwords: A survey". In *Proceedings of the Annual Computer Security Applications Conference*, Marriott University Park, Tucson, Arizona, 2005.

[14] K. Renaud and de A. Angeli. "My password is here! An investigation into visuo-spatial authentication mechanisms". *Interacting with Computers*, 16: 1017-1041, 2004.

[15] L. Sobrado and C. J. Birget. "Graphical passwords". *The Rutgers Scholar*, 4, 2002. <http://RutgersScholar.rutgers.edu/volume04/contents.htm>.

[16] F. Tari, A. A. Ozok and H. S. Holden "A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords". In *ACM International Conference Proceeding Series*, 149: 56-66, 2006.

[17] A. Perrig and D. Song. "Hash visualization: A new technique to improve real-world security". In *Proceedings of the 1999 International Workshop on Cryptographic Techniques and E-Commerce (CryTEC '99)*.

[18] Real User Corporation. "About passfaces", http://www.realuser.com/cgi-bin/ru.exe/_/homepages/technology/passfaces.htm, accessed in November 2006.

[19] D. Davis, F. Monroe and M. Reiter. "On user choice in graphical password schemes". In *Proceedings of the 13th USENIX Security Symposium*, 2004.

[20] R. Dhamija and A. Perrig. "Déjà Vu: A user study using images for authentication". In *Proceedings of the 9th USENIX Security Symposium*, 2000.

[21] A. Bauer. "Gallery of random art", 1998, <http://andrej.com/art>, accessed in December 2008.

[22] W. Jansen. "Authenticating mobile device users through image selection". *Data Security*, May 2004.

[23] Passlogix. www.passlogix.com, accessed in November 2006.

[24] E. G. Blonder. "Graphical passwords". Lucent Technologies, Inc., Murray Hill, NJ, U. S. Patent, Ed. United States, 1996.

[25] S. Wiedenbeck, J. Waters, C. J. Birget, A. Brodskiy and N. Memon. "Authentication using graphical passwords: Basic results". In *Human-Computer Interaction International (HCII 2005)*. Las Vegas, NV, 2005.

[26] S. Wiedenbeck, J. Waters, C. J. Birget, A. Brodskiy and N. Memon. "Authentication using graphical passwords: Effects of tolerance and image choice". In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. Carnegie-Mellon University, Pittsburgh, 2005.

- [27] S. Wiedenbeck, J. Waters, C.J. Birget, A. Brodskiy and N. Memon. "PassPoints: Design and longitudinal evaluation of a graphical password system". *International Journal of Human Computer Studies (Special Issue on HCI Research in Privacy and Security)*, 63: 102-127, 2005.
- [28] I. Jermyn, A. Mayer, F. Monrose, K. M. Reiter and D. A. Rubin. "The design and analysis of graphical passwords". In *Proceedings of the 8th USENIX Security Symposium*. 1999.
- [29] D. Nali and J. Thorpe. "Analysing user choice in graphical passwords". Tech. Report TR-04-01, School of Computer Science, Carleton University, Canada, 2004.
- [30] C. P. van Oorschot and J. Thorpe. "On the security of graphical password schemes". Technical Report TR-05-11. Integration and extension of USENIX Security 2004 and ACSAC 2004 papers.
- [31] J. Thorpe and P. Van Oorschot. "Graphical dictionaries and the memorable space of graphical passwords". In *Proceedings of the 13th UNIX Security Symposium*, August 2004.
- [32] J. C. Birget, D. Hong and N. Memon. "Robust discretization with an application to graphical passwords". *Cryptology ePrint Archive*, Report 2003/168, <http://eprint.iacr.org>,
- [33] K. Chalkias, A. Alexiadis and G. Stephanides. "A multi-grid graphical password scheme". In *Proceedings of the 6th International Conference on Artificial Intelligence and Digital Communications*, Thessaloniki, Greece, 2006.
- [34] A. Alexiadis, K. Chalkias and G. Stephanides. "Implementing a graphical password scheme that uses nested grids". In *Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST 2006)*, London, United Kingdom, 2006.
- [35] I. Irakleous, M. S. Furnell, S. P. Dowland and M. Papadaki. "An experimental comparison of secret-based user authentication technologies". *Information Management & Computer Security*, 10: 100-108, 2002.
- [36] B. Tribelhorn. "End user security", 2002. http://www.cs.hmc.edu/~mike/public_html/courses/security/s06/projects/index.html, accessed in November 2008.
- [37] Y. Kim and T. Kwon. "An authentication scheme based upon face recognition for the mobile environment". In *Proceedings of the International symposium on computational and information science No1*, Shanghai, China, 2004.
- [38] J. Goldberg, J. Hagman and V. Sazawal. "Doodling our way to better authentication". *CHI '02 extended abstracts on Human Factors in Computer Systems*, Minneapolis (ACM Press), 2002.
- [39] R. Weiss and A. del Luca. "PassShapes: Utilizing stroke based authentication to increase password memorability". In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*. Lund, Sweden, ACM pp. 383-392, 2008.
- [40] A. M. Eljetlawi and N. Ithnin. "Graphical password: Prototype usability survey". In *Proceedings IEEE International Conference on Advanced Computer Theory and Engineering*, pp. 351-355, 2008.

[41] K. M. Everitt, T. Bragin, J. Fogarty and T. Kohno. "A comprehensive study of frequency, interference, and training of multiple graphical passwords". In Proceedings of the 27th international conference on Human factors in computing systems. Boston, MA, USA. ACM, pp. 889-898, 2009.

[42] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot and R. Biddle. "Multiple password interference in text passwords and click-based graphical passwords". ACM CCS'09, November 9–13, 2009, Chicago, Illinois, USA, 2009.

[43] A. A. Ozok and S. Holden "A strategy for increasing user acceptance of authentication systems: Insights from an empirical study of user preferences and performance". International Journal of Business and Systems Research, 2(4): 343-364, 2008.

[44] K. Johnson and S. Werner. "Graphical user authentication: A comparative evaluation of composite scene authentication vs. three competing graphical passcode systems". Human Factors and Ergonomics Society Annual Meeting Proceedings, 52: 542-546, 2008.

[45] M. D. Hafiz, A. H. Abdullah, N. Ithnin and H. K. Mammi. "Towards identifying usability and security features of graphical password in knowledge based authentication technique". In Proceedings of the Second Asia International Conference on Modelling and Simulation, IEEE, pp. 396-403, 2008.

[46] L. Y. Por and X. T. Lin. "Multi-grid background Pass-Go". WSEAS Transactions on Information Science and Applications, 7(7): 1137-1148, 2008.

Development of Information Agent Reranking By Using Weights Measurement

Aliaa A. Youssif

*Computer Science and Information Technology/
Computer Science/Helwan University
Helwan, Egypt*

aliaay@helwan.edu.eg

Ashraf A. Darwish

*Computer Science/Mathematics/Helwan
University
Helwan, Egypt*

amodarwish@yahoo.com

Ahmed Roshdy

*Computer Science/Mathematics/Helwan
University
Helwan, Egypt*

ahmed.mind@hotmail.com

Abstract

Web search is one of the most challenging problems of the Internet today, seeking to provide users with search results most relevant to their information needs. The new improvements of search engines technologies have made available to the internet users an enormous amount of knowledge that can be accessed in many different ways. However, there are some problems that face the search engines. In this paper proposes an agent system which parses information sources and extract weights that determine the powerful of relevant information sources and prove that the word positions scores may affect on reduction the relevance of those information sources. Moreover, it will show that the user profile plays an important role in effectiveness of re-ranking and updating ranking relevant web pages where agent learns user behavior by observing user browsing for the interested result pages. In experimental work section shows how weights algorithm gets more relevant web pages than other algorithms that use word position which may reduce the value of relevance of web pages.

Keywords: Information Agent, Knowledge-base, Weights, and Re-ranking.

1. INTRODUCTION

The vast amount information source in web today rendered the intelligent information agent subtending to challenge re-ranking web pages on the fly. Given the constantly increasing information overflow of the digital age, the importance of information retrieval has become critical. Since the documents source of information retrieval have two types unstructured records and semi-structured records depending on natural language text, and also kinds of data that can be unstructured, e.g., photographic images, audio, video, etc [1,2].

The famous search engines on the marketing are now providing search facilities for databases containing billions of web pages, where queries are executed instantly [3]. But there are some problems that face search engines as following:

- The crawler-based search engines like Google can catalogue web pages and the documents automatically where it crawls the web, then the users searches through what they have found and sometimes they lead to poor queries and increase the gap between information need and request [2].
- On the other side most of search engines beside Google crawls the web pages by tracking hyperlinks to find authoritative pages using HITS algorithms that indeed get satisfied relevant pages for users queries, but they suffer from another problem that some authoritative pages have subjects and/or information sources and didn't satisfy the relevance of the user request.

Though feedback is one approach that deduces information about the user and his/her search context [4], beside that the techniques that depend on relevance feedback re-rank the search results by re-calculating the relative importance of keywords in the query.

In this paper it concentrates on the processing of information sources and those contents to extract words weights that are important to determine the content of the page subject more powerful and can re-rank web pages more efficiently [4].

The paper is organized as follows: the next section is devoted to the related work that shows the reasons of word weight importance more than its position that is used by the other agent's techniques. Section 3 contains detailed description of the proposed information agent system, showing its several components. The effectiveness of user information profile for learning agent and effecting on re-ranking web pages in sections 4 and 5, are introduced respectively. The report on an implementation and describe some experimental results in section 6 and conclusion and future work have been presented in section 7.

2. RELATED WORK

The information agents developed and have addressed many tasks, such as assisting users in searching and browsing the Web, finding, filtering and accessing large amounts of information on behalf of users, and present a reduced and potentially relevant part of this information to them [5]. Information extraction from semi-structured documents has received more attention recently. Agent in most cases is called *PersonalSearcher* [6] where it helps users to identify filtered documents with high probability of being relevant to them by transforming the search process according to users' information preferences and learns about a user's interests from the observation of user browsing on the Web.

Some techniques focus on the structure of web page like CASA [1] and Textual Case-Base Reasoning (TCBR) [7]. These algorithms depend on the position of the word either in text, line or paragraph and give a degree for its position increasing from text till paragraph which in more cases web pages puts words in text or low positions that refer to hyperlinks linked to addition information that maybe more relevant to user request. Also the advertisements web pages interest in the photos of goods and prices and little care about words positions.

Another technique in its algorithm calculate number of links that are linked to that page, and checks position level of query keywords if it is either highest, middle, or lowest in web page as in automated fuzzy agent [8]. Such this technique faces another problem that some pages have titles and keywords of user query request and have high position but the information source is not relevant for user query. Beside that some web pages have more relevant information sources and titles either in medium or low position and add some commercial advertisements in upper layer. In another hand the technique repeats itself for each page tracked by hyperlinks which make overhead in processing.

The research focuses on extracting words weights to determine the rich information, density of these words in the information source and how much they are more relevant to user query, and they effect on re-ranking web pages neither depending on them position nor subject title position.

3. THE PROPOSED AGENT ARCHITECTURE

Since the agent is an integrated part of the end-program, there are two major paradigms for the architecture. The first approach is automated treating with the results to get knowledge of certain domain and learns from user or other agent for getting feedback about user interest. The second approach is knowledge-based approach where the agent has extensive information of certain domain and learns from user profile about his/her behavior and interest to expand the feedback of the agent. The paper interests for the second approach and explain below of the architecture.

Figure 1 presents the high level view of the proposed architecture for information agent depending on words weights and information profile. The knowledge-base system which is the feedback of information agent has four major parts for processing documents as following:

- (1) Entering information sources by downloading URLs and web pages that are retrieved from the search engine or more search engines and indexing them for removing redundant URLs.
- (2) Parsing web pages by parser for extracting semi-structured data, Meta tags and link tags then determining words weights.
- (3) Re-ranking web pages by calculating relevance equation ratio.
- (4) Learning information agent from information profile about user browsing result web pages.

The user, by means of a graphical interface, submits a query to the parser by the keywords. The parser extract query to its words and find word weights for each web page which is downloaded in the knowledge-base. After that the knowledge-base calculate the relevance word weight ratio to re-ranking web pages by descending and represent re-ranked result to user interface.

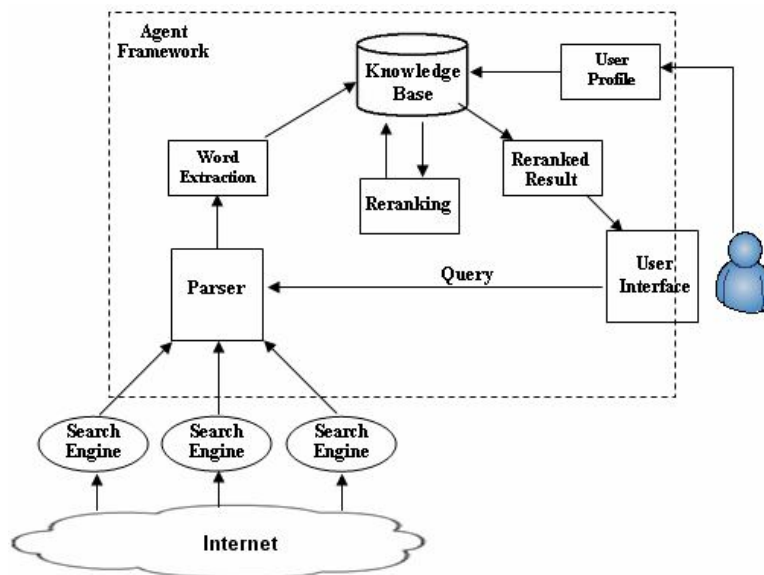


FIGURE 1: Show the high level of agent architecture

3.1. Knowledge-base

As it is explained in section 2 the importance of knowledge-base for information agent feedback in certain domain. The architecture contains on the attributes of the domain, interactions, and filtered results. The content of the base shown in Figure 2 is described as following:

DATA_EXTRACTION entity stores URL_ID (for not replication of URLs), URLs of result web pages that are downloaded from the search engine, HTML source code, meta tags, hyperlink tags, title and header tags, the number of queries that request a web page, and textfix that contains paragraphs of data source and be processed from the parser.

LOCAL_WORDS entity stores the words of each page which is extracted by the parser and those calculated weights. GENERAL_WORDS entity stores calculated word weight for all web pages that are stored in the knowledge-base.

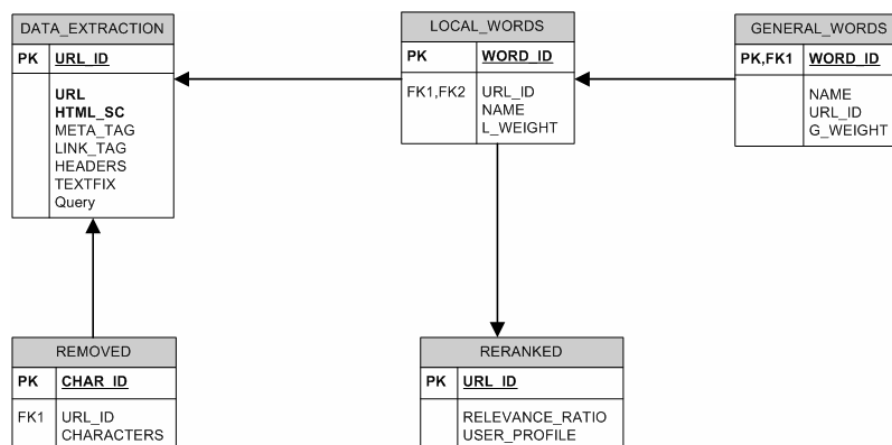


FIGURE 2: Show ER diagram of the knowledge-base

RERANKED_PAGE entity stores the calculated relevance word weight ratio for each page and counting of user browsing for web pages result. REMOVED entity stores the characters, letters, Preposition, etc that are used in paragraphs.

3.2. Parsing

The mission of the parser is extracting information sources of the web pages and determines the kind of data source according to the requirements of search engine or information agent technique [1, 9, 10, 11, 12]. The Paper focuses on the frequency of words in web pages to determine those weights and find most relevant web pages depending on the concentration of words query on those pages.

Therefore the more increasing of words query weights on a web page, the more determination of relevant web page. Example for what to discuss about, when user submit specific query request, the parser begins extract these words and compares them with those weights in knowledge-base and then finds the highly weights that refer to highly relevant web pages after calculating relevance word weight ratio.

4. USER PROFILE

Since user-profiling involves information about users of behavior-based information about them, for example their actions on browsing some pages the which learns the information agent about their interesting, besides increasing the information of the knowledge-base that are considered a feedback of user interesting [13,14].

User behavior is the most deserving sources of information for getting profiles and, from their successful interpretation; the information agent will be able to follow the user actions [15]. Therefore our agent monitors the user browsing for result pages and begins counting the user behaviors by accessing web pages; the results will be automatically updated when unobserved

voting is highly score in some web pages than others. Implicit interest indicators like time consumed in reading a Web page (considering its length), the amount of scrolling in a page, and whether it was added to the list of bookmarks or not are considered a strong correlation with explicit interest feedback, as opposed to the number of mouse clicks, which does not result in a good indicator [16,17].

The dynamic part reflects the path of user behavior information in the way of user's search. The static part lets user to show his/her desire when agent offers some options to the user. In many cases user ignores static part [18].

5. RE-RANKING

Many search engines hide the mechanism used for the document ranking this is the reason of the merging problem for the results becomes even more difficult. In addition, these kinds of approaches suffer from ignoring or knowing nothing about the user conducting the search, nor the context of the search [4].

Our evaluation for ranking of relative documents is depending on some weights using equation of relevance which is discussed in [4]. For each document d in the response to query q , the document rating is evaluated as follows (the adapted cosine function):

$$relevance_{q,d} = \frac{\sum w_d \times w_{prof} \times w_q}{W_d} \quad (1)$$

Where w_d is the weight of word in the document d , w_{prof} is the weight of chosen document d (i.e. accessing web page from different users as mentioned in section 3), we supposed its score in the first re-ranked iteration by 1, w_q is the weight of query for chosen document from agent result, and W_d is evaluated as following:

$$W_d = \sqrt{\sum (w_{prof} w_d)^2} \quad (2)$$

6. METHODOLOGY

The sequence of proposed algorithm can be obtained as following:

- 1) The user submits his/her search query.
- 2) The agent searches several other search engines with user's query.
- 3) The agent downloads the resulted web pages
- 4) The agent extracts the information source from web pages using the Parser and saves the information in the knowledge-base.
- 5) The agent calculates words weights of the query keywords.
- 6) The agent calculates relevance score for each web page and re-ranks them in a descending order.
- 7) The agent keeps track of the re-ranked web pages. Once a user browses any of these web pages, the w_{prof} score of such page increases.
- 8) Steps 1-7 will be repeated for each search.
- 9) If the search hits any of the re-ranked web pages, it will increase its w_q score.
- 10) The agent recalculates the relevance score and updates re-ranked web pages.

7. EXPERIMENTAL RESULTS

The experiment was constructed in three stages: training, retrieval and learning. For the former, the knowledge-base was built using the method proposed in section 3.1 and determined the domain to download web pages that was cars advertisements. Also parser was built to deal with

downloaded web pages for extracting information and re-ranking pages. In the retrieval stage, knowledge-base was retrieved with each first few web pages were downloaded from five search engines (Google, AltaVista, Yahoo, MSN, and Excite) and indexed these pages to prevent redundant pages. In the program we generated 20 different queries and calculated summation average of number of relevant pages and irrelevant pages. In learning stage we browsed in re-ranked pages to learn information agent of interested pages by accessing them and calculating period time of browsing for each page; the longest time taken page get score in w_{prof} and be affected for re-ranking again. The stages are described briefly as following:

7.1. Training Stage:

- (1) Build a knowledge-base and determine the domain for downloading pages.
- (2) Build parser.

7.2. Retrieval Stage:

- (1) Download first 200 web pages from each search engine.
- (2) Index web pages to prevent redundant web pages.
- (3) Generate different queries and calculated summation average.

7.3. Learning Stage:

For each term, Browsed in re-ranked pages to give interested pages scores in w_{prof} for next re-ranking.

The Figure 4 represents the most common words of user query that insure the relevance of web page.

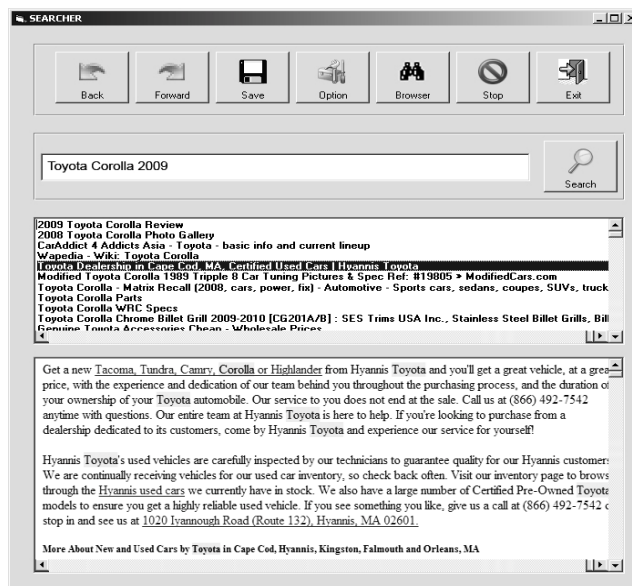


FIGURE 3: Show Agent Search Result

7.4. Comparative Analysis

In order to evaluate our proposed algorithm, we compared our findings with the score resulted from TCBR [7], CASA [1] Automated fuzzy algorithm [8]. Table 1 shows the comparative evaluation of common queries that have effect of word position in average of first 20 relevant web pages.

| | Automated fuzzy score | TCBR | CASA score | Proposed score |
|--------------------|-----------------------|-------|------------|----------------|
| MAZDA ph | 3.016 | 2.255 | 0.918 | 4.565 |
| FIAT Audio | 3.927 | 3.573 | 1.641 | 5.126 |
| BMW Engine | 4.537 | 3.847 | 2.177 | 6.337 |
| HONDA in Boston | 2.472 | 1.501 | 0.861 | 3.691 |
| Toyota Accessories | 4.582 | 3.674 | 2.255 | 5.969 |
| PROTON-GN2 Weight | 3.415 | 2.293 | 1.106 | 4.846 |
| PEUGEOT Spares | 4.357 | 3.063 | 2.144 | 5.138 |
| JEEP Cylinder | 2.732 | 1.587 | 0.161 | 3.813 |
| CHERRY | 3.157 | 2.059 | 0.345 | 4.152 |
| NISSAN Carpet | 3.437 | 2.462 | 1.147 | 4.598 |
| Used LANOS | 3.843 | 2.826 | 1.164 | 5.326 |
| Mercedes Brake | 2.952 | 2.147 | 1.030 | 4.293 |
| SEAT Filter | 2.638 | 1.476 | 0.185 | 3.232 |
| Hyundai Lights | 2.431 | 1.630 | 0.289 | 3.813 |
| SKODA Hatchback | 4.362 | 3.326 | 2.577 | 6.429 |

TABLE 1: The difference relevant scores between algorithms

From Table 1, we noticed that our proposed algorithm gives higher word score than CASA, TCBR and fuzzy algorithm. In computing the relevance of a web page, our equation does not take into account the word location.

However, in some queries like “Toyota Accessories”, “Proton-GN2 Weight”, “Honda in Boston” the factor of word location in CASA algorithm and Δ_j in TCBR equation for word weight $weight(w_p, d_i) = tf_{ij} + \Delta_j$ decrease (where tf_{ij} is term frequency that determines the frequency of word appearance in document) because these words occurred more frequently in lines than in paragraphs and titles.

Moreover, in the automated fuzzy algorithm we noticed that the Position_Score of some words decreases to some extent. This is due to the fact that these words occurred in the middle of the web pages, which affected their score. The relevant web page score is determined by the equation:

$$Score = (2 * Frequency_Score) + Position_Score + Links_Score$$

Summary:

Figure 3 shows the comparison results between 3 algorithms: CASA algorithm [1], TCBR algorithm [7], automated fuzzy algorithm [8] and weights measures algorithm of the first 200 re-ranked web pages from information agent and that describes the performance of agent can be more efficient depending on weights measurement.

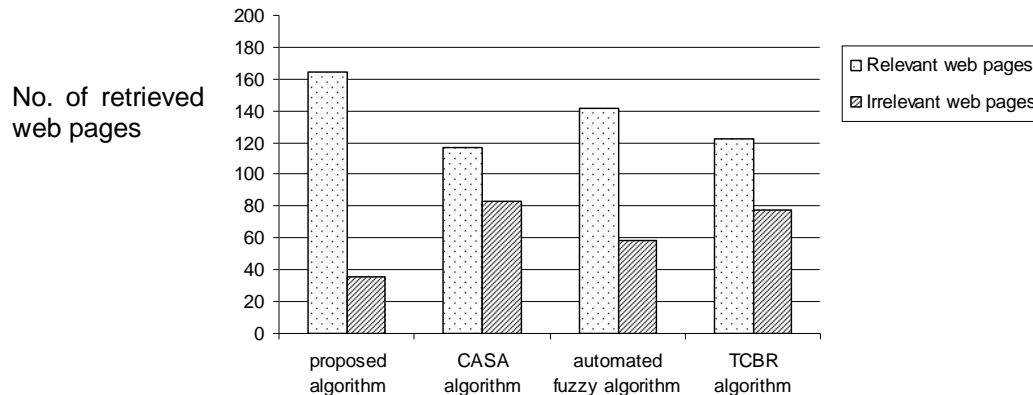


FIGURE 4: Comparison between four algorithms

The performance of information agent can be increased by improving the mechanism of downloading URLs and web pages using semantic web techniques as in [11,12,19] that determine word vectors of context meaning , and reaching to suitable solution to the retrieve relevant documents dynamically.

8. CONCLUSION AND FUTURE WORK

In this paper we scoped on filtering powerful information source by parsing the information to extract weights for finding more relevant pages and re-ranking web pages with learning information agent of user profile about interesting web pages that can be more efficient in re-ranking. The result of the experiment for proposed algorithm is 0.82% of relevant web pages compared to highly percentage of other algorithms about 0.71%. We used Microsoft Access to build knowledge-base and Visual Basic to build the other components of information agent.

9. REFERENCES

- [1] X. Gao, "A methodology for building information agents", in Yang, Y., Li, M., Ellis, A. (Eds), *Web Technologies and Applications*, International Academic Publishers, <<http://www.cs.mu.z.au/~xga/apweb/index.htm> >, pp.43-52. (1998)
- [2] Ed Greengrass "Information Retrieval: A Survey", Available at: <http://www.csee.umbc.edu>. (2000)
- [3] M. Caramia, G. Felici, and A. Pezzoli "Improving search results with data mining in a thematic search engine", *Computers and Operations Research* Volume 31, Pages: 2387 – 2404, ISSN: 0305-0548. (2004)
- [4] B. Chidlovskii, N.S. Glance and M.A. Grasso, "Collaborative Re-Ranking of Search Results", Available at: <http://citeseer.ist.psu.edu>. (2000)
- [5] Daniela Godoy, Silvia Schiaffino, Analia Amandi "Interface agents personalizing Web-based tasks", *Cognitive Systems Research* Volume 5, Issue 3, Pages 207-222. (2004)
- [6] Godoy, D., & Amandi, A. "Personalsearcher: An intelligent agent for searching web pages" In *International joint conference IBERAMIASBIA'2000* (pp. 43–52). Atibaia, Sao Paulo, Brazil: Springer. (2000)
- [7] D. Lis Godoy, "Generating User Profiles for Information Agents", Available at: <http://citeseer.ist.psu.edu>. (2003)
- [8] M. Mohammadian in the book "Intelligent Agents for Data Mining and Information Retrieval", Idea Group Publishing (an imprint of Idea Group Inc.), ISBN: 1-59140-277-8. (2004)
- [9] L. Shen and A.K. Joshi, "An SVM Based Voting Algorithm with Application to Parse Reranking", the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Pages: 9 - 16. (2003)
- [10] A. Penev and R. Wong, "Shallow NLP techniques for Internet Search", the 29th Australasian Computer Science Conference - Volume 48, Pages: 167 – 176, ISBN ~ ISSN: 1445-1336, 1-920682-30-9. (2006)

- [11] C. Cesarano, A. d'Acierno and A. Picariello, "An Intelligent Search Agent System for Semantic Information Retrieval on the Internet", the 5th ACM international workshop on Web information and data management, Pages: 111 – 117, ISBN:1-58113-725-7. (2003)
- [12] G. Wisniewski & P. Gallinari "From Layout to Semantic: A Reranking Model for Mapping Web Documents to Mediated XML Representations" proceeding of RIAOCID (2007) Conference, URL: <http://dblp.uni-trier.de/db/conf/riao2007.html#WisniewskiG07>. (2007)
- [13] Bas van Gils and Eric D., "User-profiles for Information Retrieval", Methodologies for Intelligent Systems, volume 542 pages 102-111 ISBN: 978-3-540-54563-7. (2006)
- [14] S.E. Middleton, "Capturing knowledge of user preferences with recommender systems", Available at: <http://citeseer.ist.psu.edu>. (2003)
- [15] D. Godoy and A. Amandi "User profiling in personal information agents: a survey", The Knowledge Engineering Review, Volume 20, Pages: 329 - 361 ISSN: 0269-8889. (2005)
- [16] Claypool, M., Le, P., Wased, M., & Brown, D. "Implicit interest indicators. Intelligent User Interfaces", Proceedings of the 6th international conference on Intelligent user interfaces Santa Fe, New Mexico, United States Pages: 33 – 40 ISBN:1-58113-325-1 (2001)
- [17] Kelly, D., & Belkin, N. J. "Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevant feedback", proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (pp. 408–409). New Orleans, LA, USA: ACM Press. (2001)
- [18] C. Dharap "Context-based and user-profile driven information retrieval", Fremont CA (US), Philips Corporation New York (2001).
- [19] G. Cheng, W. Ge and H. Wu, "Searching Semantic Web Objects Based on Class Hierarchies", proceeding of LDOW Conference, available at: <http://iws.seu.edu.cn/publications/cgwwq08> . (2008)

COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA