David C. Wyld
Jan Zizka (Eds)

Computer Science & Information Technology

Sixth International Conference on Computer Science, Engineering and Applications (CCSEA 2016)
Dubai, UAE, January 23~24, 2016

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-46-5
DOI : 10.5121/csit.2016.60201 - 10.5121/csit.2016.60217

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Sixth International Conference on Computer Science, Engineering and Applications (CCSEA-2016) was held in Dubai, UAE, during January 23~24, 2016. The Fifth International Conference on Cloud Computing: Services and Architecture (CLOUD-2016), The Fourth International Conference on Data Mining & Knowledge Management Process (DKMP-2016), The Fifth International Conference on Software Engineering and Applications (SEA-2016) and The Second International Conference on Signal and Image Processing (SIPRO-2016) were collocated with the CCSEA-2016. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCSEA-2016, CLOUD-2016, DKMP-2016, SEA-2016, SIPRO-2016 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CCSEA-2016, CLOUD-2016, DKMP-2016, SEA-2016, SIPRO-2016 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCSEA-2016, CLOUD-2016, DKMP-2016, SEA-2016, SIPRO-2016.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld
Jan Zizka

Organization

General Chair

Jan Zizka	Mendel University in Brno, Czech Republic
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia

Program Committee Members

Abd El-Aziz Ahmed	Cairo University, Egypt
Abdallah Rhattoy	Moulay Ismail University, Morocco
Abdolreza Hatamlou	Islamic Azad University, Iran
Adam Przybylek	Gdansk University of Technology, Poland
Adnan H. Ali	Institute of Technolgy, Iraq
Ali Elkateeb	University of Michigan-Dearborn, USA
Ali El-Zaart	Beirut Arab University, Lebanon
Ali Hussein Mohammed	Alexandria University, Egypt
Ali Zaart	Beirut Arab University, Lebanon
Al-Majeed	University of Essex, UK
Ankit Chaudhary	Truman State University, USA
Arif Sari	Girne American University, Cyprus
Ayad Salhieh	Australian College of Kuwait, Kuwait
Baghdad ATMANI	University of Oran, Algeria
Barkat Warda	University of Constantine 1, Algeria
Chikh Mohammed Amine	University of Tlemcen, Algeria
Chiranjib Sur	University of Florida, US
Dongchen Li	Peking University, China
Faiz ul haque Zeya	Bahria University, Pakistan
Farzad Kiani	Istanbul S.Zaim University, Turkey
Hossein Jadidoleslamy	MUT University, Iran
Ing. Habil. Natasa Zivic	University of Siegen, Germany
Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University, Egypt
Israashaker Alani	Ministry of Science and Technology, Iraq
Izzat Alsmadi	Damascus University, Syria
Jalel Akaichi	University of Tunis, Tunisia
Jan Lindström	MariaDB Corporation, Finland
Jan Zizka	Mendel University in Brno, Czech Republic
Japhynth J	Dr.G.U.Pope College of Engineering, India
José Raniery	University of São Paulo, Brazil
José Vargas-Hernández	University of Guadalajara, Mexico
Laura Felice	Universidad Nacional del Centro, Argentina
Li Zheng	University of Bridgeport, USA
Liyakath Unisa	Prince Sultan University, Saudi Arabia

Madhavi Vaidya	Mumbai university, India
Mahdi Mazinani	University of Tehran, Iran
Mahfuzul Huda	Integral University, India
Manoj Vasanth Ram	Lead Hardware Engineer, USA
Manu Sood	Himachal Pradesh University, India
Maryam Rastgarpour	Islamic Azad University, Iran
Marystella Amaldas	Saigon International University, Vietnam
Meachikh	University of Tlemcen, Algeria
Mehrdad Jalali	Mashhad Azad University, Iran
Mohamed Sahbi Bellamine	University of Carthage, Tunisia
Mohammadreza Balouchestani	University of New Haven, USA
Mohammed AbouBakr Elashiri	Beni Suef University, Egypt
Mohammed Amin	Higher Colleges of Technology, UAE
Moses Ekpenyong	University of Uyo, Nigeria
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Mustafa Salah	University Putra Malaysia, Malaysia
Natarajan Meghanathan	Jackson State University, USA
Nishant Doshi	Marwadi Education Foundation, India
Noria Benyettou	Ecole National Polytechnique of Oran, Algeria
Noureddine Hassini	University of Oran, Algeria
Rafa E.Al-Qutaish	University of Quebec, Canada
Rafah M. Almuttairi	University of Babylon, Iraq
Rahil Hosseini	Islamic Azad University, Iran
Ramayah T	Universiti Sains Malaysia, Malaysia
Ramesh S	Dr Ambedkar Institute of Technology, India
Rastgarpour M	Science and Research University, Iran
Revathi V	Adhiyamaan College of Engineering, India
Reza Ebrahimi Atani	University of Guilan, Iran
Rim Haddad	Innov'com Laboratory, Tunisia
Ritambhra Korpal	University of Pune, India
Rosziati Ibrahim	Universiti Tun Hussein Onn Malaysia, Malaysia
Saad M. Darwish Saad	Alexandria University, Egypt
Saeid Asgari Taghanaki	Azad University, Iran
Salah Al-Majeed	Military Technological College, Oman
Samadhiya	National Chiao Tung University, Taiwan
Sandhya	Gautam Buddha University, India
Sattar B. Sadkhan	University of Babylon, Iraq
Seyed Ziaeddin Alborzi	Université de Lorraine, France
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Shakir Khan	Leading University, Bangladesh
Souad Bekkouche	Djilalli Liabbes University, Algeria
Soumen Kanrar	Vehere Interactive Pvt Ltd, India
Stefano Berretti	University of Florence, Italy
Subarna Shakya	Tribhuvan University, Nepal
Tanweer Alam	Islamic University, Kingdom of Saudia Arabia
Vasanth Mehta	SCSVMV University, India
Venkata Raghavendra	Adama University, Ethiopia
Yahya Slimani	University of Manouba, Tunisia
Zhang Xiaojun	Dublin City University, Ireland

Technically Sponsored by

Networks & Communications Community (NCC)

Computer Science & Information Technology Community (CSITC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Sixth International Conference on Computer Science, Engineering and Applications (CCSEA 2016)

The Fifth International Conference on Cloud Computing: Services and Architecture (CLOUD 2016)

The Fourth International Conference on Data Mining & Knowledge Management Process (DKMP 2016)

The Fifth International Conference on Software Engineering and Applications (SEA 2016)

The Second International Conference on Signal and Image Processing (SIPRO 2016)

EFFICIENT CALL PATH DETECTION FOR ANDROID-OS SIZE OF HUGE SOURCE CODE

Koji Yamamoto and Taka Matsutsuka

Fujitsu Laboratories Ltd., Kanagawa, Japan
{yamamoto.kouji,markn}@jp.fujitsu.com

ABSTRACT

Today most developers utilize source code written by other parties. Because the code is modified frequently, the developers need to grasp the impact of the modification repeatedly. A call graph and especially its special type, a call path, help the developers comprehend the modification. Source code written by other parties, however, becomes too huge to be held in memory in the form of parsed data for a call graph or path. This paper offers a bidirectional search algorithm for a call graph of too huge amount of source code to store all parse results of the code in memory. It refers to a method definition in source code corresponding to the visited node in the call graph. The significant feature of the algorithm is the referenced information is used not in order to select a prioritized node to visit next but in order to select a node to postpone visiting. It reduces path extraction time by 8% for a case in which ordinary path search algorithms do not reduce the time.

KEYWORDS

Call graph, Graph path, Bidirectional search, Static source code analysis, Huge amount of source code

1. INTRODUCTION

Source code written by other parties, especially open source code, are often utilized to build developers' own software products and services.The developers merge other party's code as a library, or adds their own original functions into it in order to make their software high value with competitive development cost.In the latter case, it is the key to success to understand the overlook and details of the source code.

It is an effective approach of source code comprehension to recognize relationships between classes or methods in the code. The main kinds of relationships for imperative object oriented programming languages like Java and C++ are caller-callee relationship (as known as call graph [1]), data structure, and class inheritance.We believe grasping caller-callee relationship, especially a caller-callee relationship path (abbreviated "call path" hereafter), is one of the best entries to comprehend source code because it highlights outlook of behaviours of the executed code so as to emphasize which methods/classes have to be first investigated in detail. Call paths are acquired using static program analysis.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 01–12, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60201

The size of open source code is increasing huge. For instance, open source version of Android OS source [2] consists of 50 to 100 million lines of code. In spite of that, the required time for static analysis of source code has been reduced drastically today. Understand ™ [3] by Scientific Toolworks, Inc., for example, consumes less than one tenth of static analysis time of the previous tools such as Doxygen [4] in our experience. Static analysis of huge scale source code written by other parties is now the realistic first step to comprehend them if the following problem is resolved.

Huge amount of analysis result of huge size source code is, however, still barrier to understand the source code for an ordinary development environment. A developer usually has a general type of laptop/desktop computer with tiny memory, at most 16GB. The memory does not store all the result if the target source code is for Android OS, 50-100 million lines of code or similar size of code. Actually a server with much more size of memory cannot treat the result efficiently. It takes much more time to extract a call path. It disturbs developers' source code comprehension.

Our contribution is a bidirectional search algorithm to extract a call path from a call graph of too huge source code to store all parse results of the code in memory. It reduces 8% of path extraction time for a case in which ordinary path search algorithms do not reduce the time.The first characteristic feature of the algorithm is it refers to a method definition in source code corresponding to the visited node in the call graph. The second significant feature is the referenced information is used not in order to select a prioritized node to visit next but so as to select a node to “postpone” visiting. They are dedicated to the search time reduction. In addition, the algorithm halves the required time for the aforementioned case if all the data is stored in memory, though it is far from a real situation.

In the rest of this paper, we explain call graphs themselves and graph search algorithms. After that, we introduce our bidirectional search algorithm for call graphs and its evaluations. Then we make some discussions followed by concluding remarks.

2. DEFINITIONS

2.1. Call Graph

A call graph is the directed graph where the nodes of the graph are the methods (in Java; the functions in C++) of the classes of the program; each edge represents one or more invocations of a method by another method [1]. The former method is referred to as a callee method, and the latter as a caller method. The staring node of the edge corresponds to the caller method. The ending node stands for the callee method.

An example of call graphs is depicted in Figure 1, which is retrieved from the source code shown in Figure 2. Each bold text part in Figure 2 corresponds to the node having the same text in Figure 1. Information on methods and their invocations is retrieved from source code using static analysis tools like as [3] and [4]. In practice graph could be more complicated: Each method might be invoked by other methods than method “transmit()”; Method “transmit()” itself might be invoked by some methods.

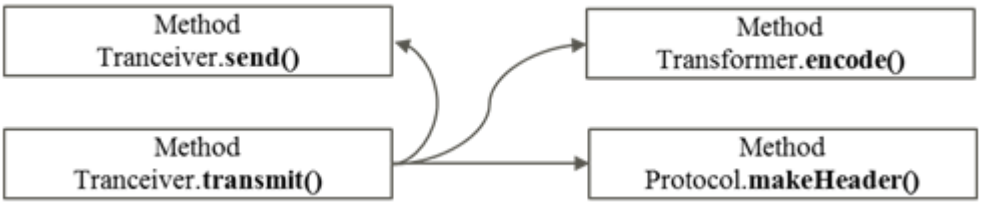


Figure 1. An example of call graphs

```
public class Tranceiver implements ITranceiver {
    private Transformer transformer;
    private Protocol protocol;
    private boolean send(Host destination) { ... }
    ...
    public boolean transmit(String data, Host destination) {
            byte[] encodedData = transformer.encode(data);
            header = protocol.makeHeader();
            send(header, encodedData, destination);
    }
}
```

Figure 2. Source code corresponding to the call graph in Figure 1

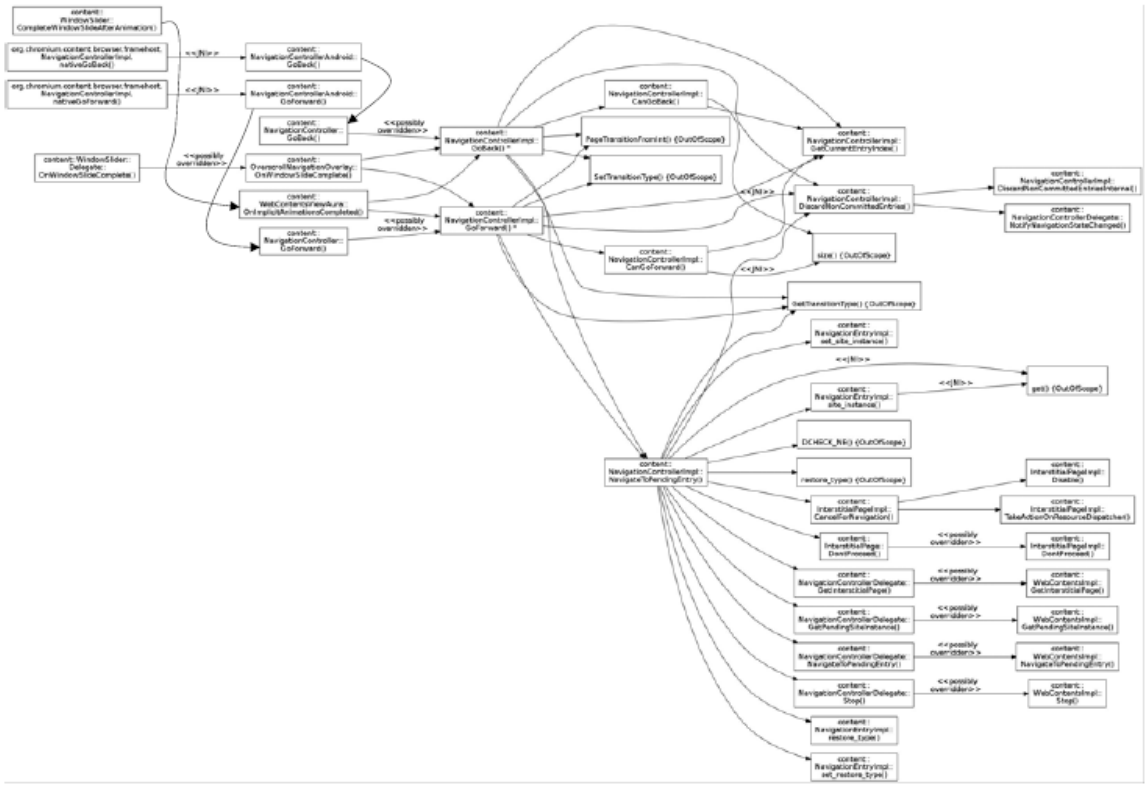


Figure 3. A complicated example of call graphs

2.2. Call Path

A call path is a sequence of edges where all the edges are connected and each edge is connected to at most one incoming edge and at most one outgoing edge. The edge having no incoming edge is called as an “initial node.” The edge with no outgoing edge is called as a “final node.” The methods corresponding to the initial node and the final node are called as an “initial method” and a “final method” respectively.

Figure 4 shows an example of call paths, which is extracted from a little bit complicated call graph shown in Figure 3. In most cases, extracted call paths are simpler to grasp caller-callee relationship than general call graphs for developers if they know the names of an initial method and a final method to be concerned.

Figure 4. An example of call paths, which is extracted from the graph in Figure 3

2.3. Bidirectional Search Algorithm for Call Graphs

A bidirectional graph search algorithm is a graph search algorithm that finds a shortest path from an initial node to a final node in a directed graph.It traverses nodes forward in the graph and traverses nodes backward simultaneously [5]. Recent bidirectional algorithms use a heuristic distance estimate function to select a node to visit next [6] [7] [8]. An example of heuristic functions is Euclidean distance of a pair of nodes for the corresponding real distance is Manhattan distance.

To our knowledge, no heuristic estimate function for a call graph has been found yet.

3. BIDIRECTIONAL SEARCH ALGORITHMS FOR CALL GRAPH

We present an algorithm for bidirectional search in a call graph. The algorithm use source code properties corresponding to a visiting node in order to select a next node to visit, while other bidirectional search algorithms use a heuristic estimate value between a visiting node and the initial/final node [6] or frontier nodes [7][8]. In our algorithm 'bidir_postpone', a next visiting node is decided using the type of method corresponding to the visiting node in addition to the outgoing or incoming degree (outdegree or indegree) of the node.

```
procedure bidir_postpone(E, V, initialNode,
                                    finalNode):
 1: todoF := {initialNode}
 2: todoB := {finalNode}
 3: /* Let prevF and prevB be dictionaries
          of type V to V.
       Let delay, distF, and distB be dictionaries
          of type V to integer. */
 4: for v in V do
27:       if frwd then
28:          next := {v | (u->v) in E}
29:       else
30:          next := {v | (v->u) in E}
31:       endif
32:       for node v in next; do
33:          alt := dist[u] + 1 /* all the edge
                       weighs one in a call graph.
    */
```

```
 5:    prevF[v] := None; prevB[v] := None
 6:    delay[v] := 0
 7:    distF[v] := ∞; distB[v] := ∞ /* which
           stand for the distance from v along
   with
           forward (F)/backward (B) path. */
 8: endfor
 9: intermed := None
10: while |todoB| > 0 or |todoF| > 0 do
11:    if |todoB| < |todoF| then
12:       frwd := True; todo := todoF;
              frontiers := todoB; dist := distF
13:    else
14:       frwd := False; todo := todoB;
              frontiers := todoF; dist := distB
15:    endif
16:    todo2 := { }
17:    for node u in todo; do
18:       if delay[u] > 0 then
19:          delay[u] := delay[u] - 1
20:          add u to todo2
21:          continue for-loop with next node u
22:       else if the type of the class of
                        the method corresponding to
   u
                        is interface and not frwd
   then
                        not frwd then
23:          add u to todo2
24:          delay[u] := 3 - 1
25:          continue for-loop with next node u
26:       endif
34:          if dist[v] > alt then
35:             prev[v] := u
36:             dist[v] := alt
37:             if v in frontiers then
38:                intermed := v
39:                exit from while
40:             endif
41:             add v to todo2
42:          endif
43:       endfor
44:       if frwd then todoF := todo2
45:       else todoB := todo2 endif
46:       todo2 := { }
47:    endfor
48: endwhile
49: if intermed is None then
50:    output ERROR
51: else
52:    v := intermid
53:    while prevF[v] is not None do
54:       output (prevF[v] -> v)
              as a path constituent
55:    endwhile
56:    v = prevB[intermid]
57:    while prevB[v] is not None do
58:       output (v -> prevB[v])
              as a path constituent
59:    endwhile
60: endif
```

Figure 5. Algorithm 'bidir_postpone'

Another difference between our algorithm and the previous algorithms is that the selected node by our algorithm is not a prioritized node to visit next but a node that is high cost to visit. The node will be scheduled to visit some steps later instead of visiting immediately.

Figure 5 shows our algorithm 'bidir_postpone.' E and V in parameters in the figure stand for a set of edges and a set of nodes in the call graph respectively.

At line 22, the algorithm determines next node to visit should be treated immediately or should be postponed treating. If the corresponding class type of the visiting node is interface in Java or abstract class in C++, the treatment is postponed. This type of method can be called by many methods. Therefore indegree of the corresponding node is greater than usual nodes. That is why visiting to such nodes should be postponed. If the visiting node is the case, treatment of the node will be suspended for 3 steps (See the line 24 and lines 18-20).

It is false that the postponement can be achieved by assigning heavy weight to edges adjacent to the nodes that hold the condition at line 22, instead of the treatment suspension because an algorithm using such heavy edges may output longer path than algorithm bidir_postpone.

4. EVALUATION

We compare the results of applying four types of algorithms to four pairs of initial and terminal nodes (hereupon the both nodes are referred to as starting point nodes) under two kinds of conditions. The result tells (1) if all the data is stored in memory, our algorithm reduces the duration for the significant case where the original bidirectional search does not reduce time compared to more naïve unidirectional search. (2) Even if the data is stored in HDD, our algorithm reduces the path extraction time by 8% for the aforementioned case. The details are shown hereafter.

4.1. Algorithms, data, and conditions to compare

The algorithms are the following four types. The second and third ones are almost the same as our algorithm itself:

A1 ‘Bidir_3postpone’: It is the algorithm shown in Figure 5.

A2 ‘Bidir_6postpone’: It is slightly modified version of algorithm of A1 ‘Bidir_3postpone’ with delay 6. It delays node visiting for 6 steps at line 24 in Figure 5 instead of 3 steps. The purpose to compare the algorithm A1 with A2 is to check whether postponement step of algorithm A1 is adequate or not.

A3 ‘Bidir_0postpone’: It is almost the same algorithm as A1 ‘Bidir_3postpone’ and A2 ‘Bidir_6postpone.’ In this algorithm, the condition at line 22 in Figure 5 is always false while the property in the source code, which is the type of the corresponding method, is retrieved from the parse result stored in HDD. The algorithm is for evaluation of an overhead of the parse result retrieval.

A4 ‘Bidir_balanced’: It is the almost original version (appeared in [9]) of bidirectional search algorithm with no heuristic estimate functions, due to missing of estimate functions for call graphs. The difference between the algorithm and A1 in Figure 5 is that lines 18 through 26 are omitted and the rest of the for-loop starting from line 32 is always executed.

The starting point node pairs are as follows:

P1 ‘A->C’: The number of reachable nodes by traversing forward from the initial node is much more than the number of reachable nodes by traversing backward from the final node.

P2 ‘C->N’: The numbers of forward nodes from the initial node and backward nodes from the final node are both few.

P3 ‘N->R’: Opposite pattern of P1 ‘A->C’. The number of forward nodes from the initial node is much less than backward nodes from the final node.

P4 ‘A->R’: The numbers of forward nodes from the initial node and backward nodes from the final node are both many.

Figure 6 shows the numbers of nodes reachable from the initial node and the final node. For P2 and P4, the numbers of both type of nodes are almost the same. They are different from each other for P1 and P3. Note that the measurement of the number is logarithmic. For instance, the number for the final node for P3 is 5 times greater than one for the initial node.

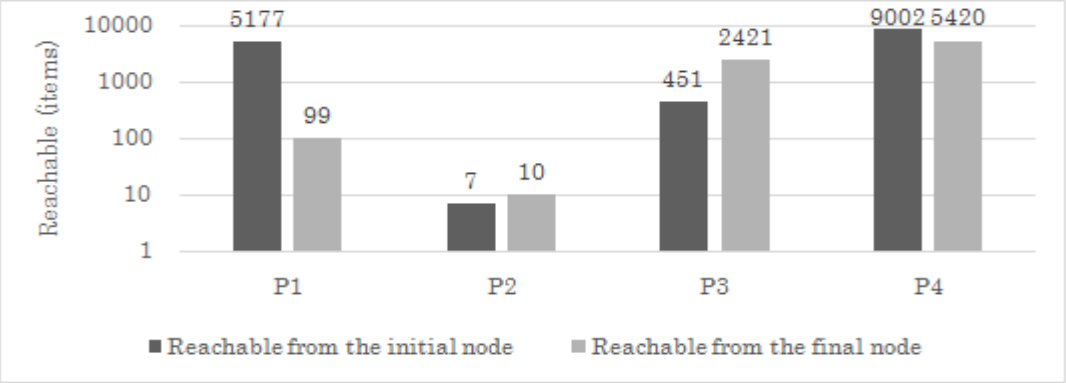


Figure 6. The numbers of nodes reachable from starting points

The conditions are the following two patterns. The latter is more similar to practical use case:

C1 ‘In memory’: All graph data is stored in memory.

C2 ‘In HDD’: A part of graph data is constructed on demand from the source code parse results that are stored in hard disk drive. Some graph edges and all graph nodes correspond to some parse result straightforwardly. Some other edges should be built up with syntactical analysis results and lexical analysis results.

C1 is the condition to measure the performance of the algorithms themselves. C2 is the condition that is almost the same as the actual execution environment with an exception. The environment contains parse result of source code stored in hard disk drive, and a result extraction program to convert specified part of the parse result to graph edges and nodes. The exception that differs from the actual environment is all the data stored in hard disk drive is not cached into memory (that is disk cache) at initial time. It makes measurement variance due to disk cache very small.

The source code for evaluation from which the call graph is constructed is partial source files of practical source code of Android OS for smartphones, version 4.4.4_r2 in [2]. The number of methods in the partial source files set is about 2% of the number in the whole source set. The parsed data for only 2% of the whole code occupies 2 GBytes size or more. For ordinary lap top computers of developers, even this size of partial source files are too huge to be held in memory.

4.2 Results

The measurements are executed 3 times. The average measured values are described in Figures 7 to 9, where logarithmic scale is used for all the Y axes. Figure 7 shows the time to traverse under

the condition C1 'In memory'. Figure 8 is for the time to traverse under the condition C2 'In HDD'. Figure 9 tells the number of visited nodes. Note that an I-shaped mark at the top of each plotted box, in Figures 7 and 8, stands for the range of the sample standard deviation, $+\sigma$ and $-\sigma$. Three times measurements seem enough because the deviations are sufficiently small.

Figure 7 depicts algorithm A1 'Bidir_3postpone' halves the time to traverse for the case in which the numbers of reachable nodes from the starting point nodes are both large (P4). The precise ratio is 1.07 seconds for our algorithm (A1) to 2.71 seconds for the original algorithm (A4), which stands for 60.5% reduction. Actually in other cases P1, P2, and P3, naïve bidirectional search (A4) runs in much shorter time than a unidirectional search, which is the directed edge version of Dijkstra algorithm, and is much more naïve than A4. Thus the time reduction in the case P4 is most desired by developers.

The resulting time to traverse for in-memory access case (C1, in Figure 7) is almost proportional to the number of nodes to be visited by each algorithm, shown in Figure 9. Therefore the less nodes are visited by the algorithm, the less traversal time can be achieved.

In contrast to the in-memory access, the results of the cases in which the data is stored in HDD (C2), in Figure 8, tell our algorithm takes worse time than the original algorithm (A4) for the case P3. In the case P4 that is most desired to reduce the time by developers, however, our algorithm (A1) spends 281.9 seconds and the original algorithm (A4) consumes 307.8 seconds. The difference is 25.9 seconds, which means ours (A1) achieves 8.4% reduction to the original (A4).

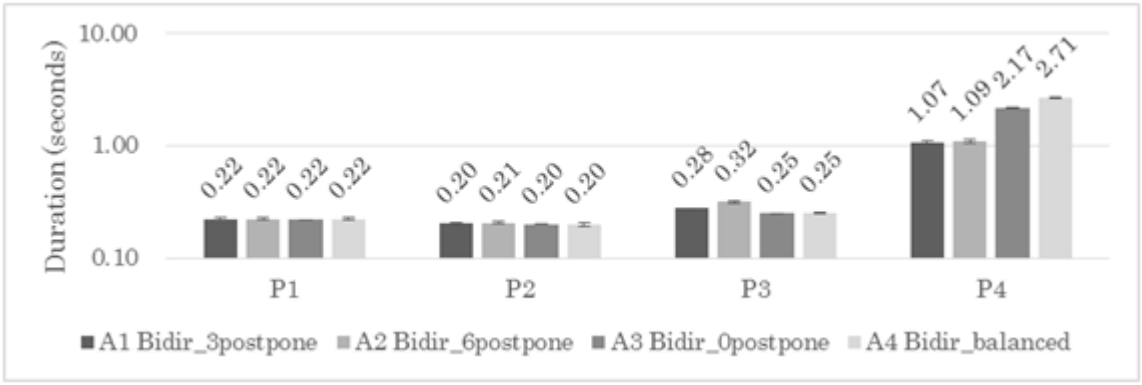


Figure 7. Time to traverse (in memory)

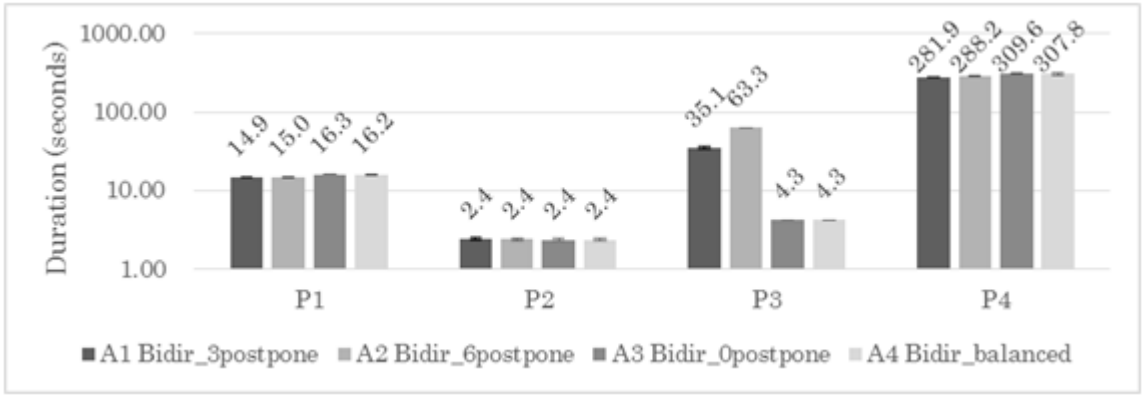


Figure 8. Time to traverse (in HDD)

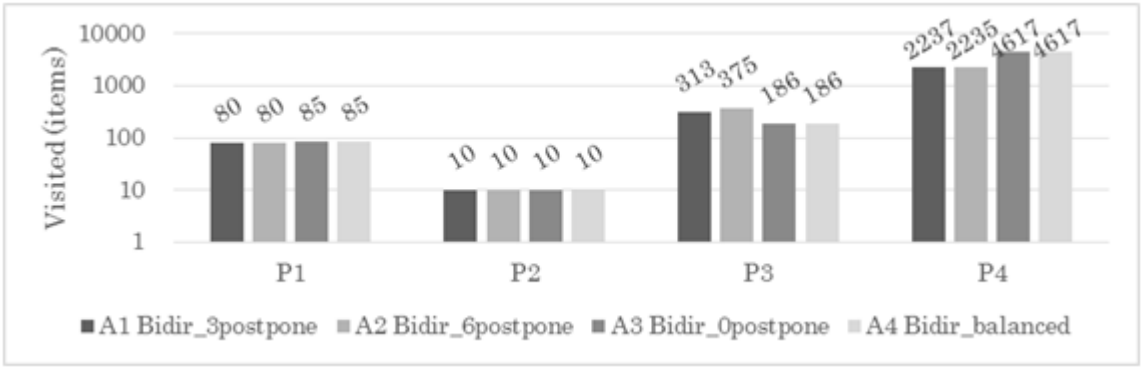


Figure 9. The numbers of visited nodes

4.3. Barriers to time reduction

We assume the overhead time for in-HDD case (C2, in Figure 8) comes from extra disk accesses to retrieve the properties of methods that occur at line 22 in Figure 5. Our disk access method in our algorithm is still naïve. Therefore the effect of the disk accesses could be reduced by the result of further investigation.

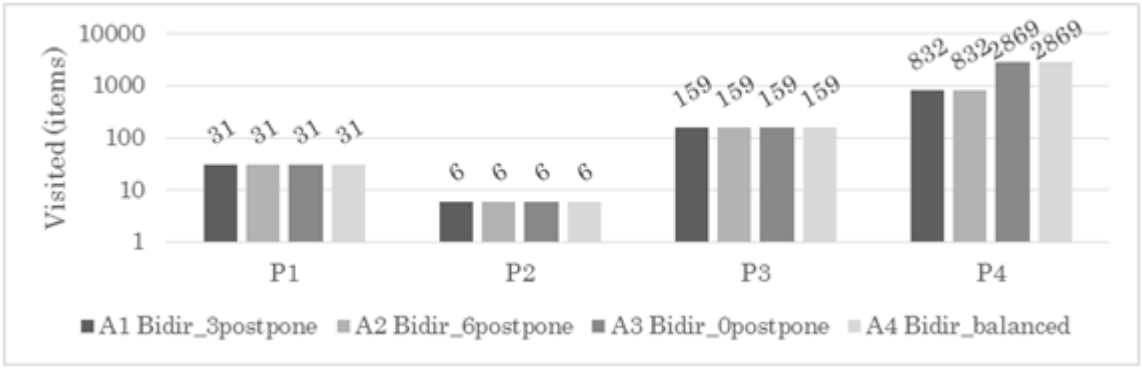


Figure 10. The numbers of nodes that the algorithms visited forward

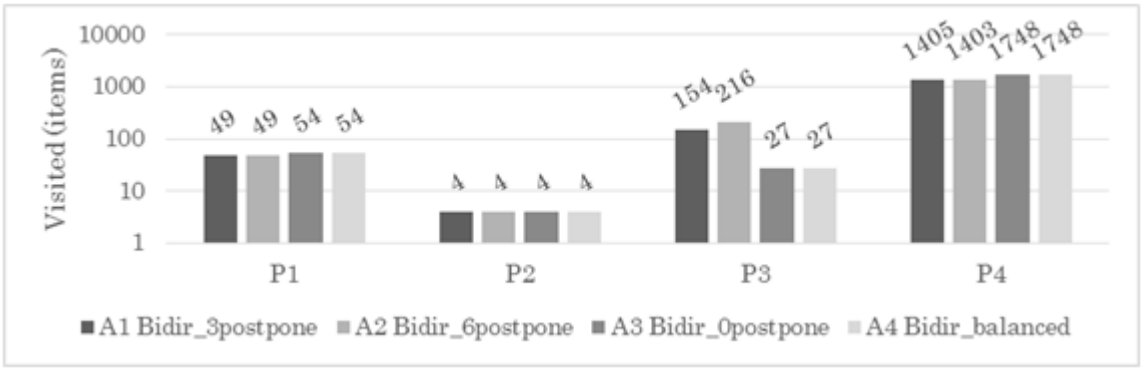


Figure 11. The numbers of nodes the the algorithms visited backward

The algorithm A1 visits remarkably large number of nodes for the starting point pair P3, in which the number of forward edges reachable from the initial node is much less than backward edges

reachable from the final node. The algorithm traverses forward the same number of nodes from the initial node of P3 as the other algorithms (See Figure 10). It visits backward five times as large number of nodes from the final node of P3 as the algorithms A3 and A4, as shown in Figure 11. In the both algorithms A1 and A4, the forward search and the backward search meet at the same node. Note that the meeting point node is expressed as 'intermed' at line 38 in Figure 5. The node adjacent to the meeting point node (abbreviated as 'MP node' hereafter) in forward direction holds the postponement condition described at the line 22 in Figure 5. The adjacent node is on the path between the MP node and the final node. Thus visiting backward to the MP node is postponed for 3 steps. One hundred and twenty seven extra nodes has been visited while that. P3 notices our algorithm can be improved using other kinds of information than the type of the class corresponding to the visited node.

5. DISCUSSION

5.1. Intermediate Nodes

Actually resulting paths for our evaluation data P1 through P3 can be concatenated because the final node for Pn is the initial node for P(n+1) for n=1, 2. In addition, the concatenated path starts from the initial node of P4 and ends with the final node of P4.

It is interesting the summation of the numbers of visited nodes for P1 through P3 is much less than the numbers of visited nodes for P4 (See Figure 9.) It states the possibility of existence of somewhat “efficient” intermediate nodes for path search of source code call graph. If as bidirectional search algorithm can start traversing from the “efficient” intermediate nodes in addition to starting point nodes, it takes less time to extract call path than our current version of algorithm. We will try to find the feature of such type of intermediate nodes.

Note: The resulting path for P4, however, does not include the initial node and the final node of P2. Therefore we treat P1 through P4 as almost independent example data from each other.

5.2. More Aggressive Use of Features in Source Code

Our algorithm shown in Figure 5 only uses the class type corresponding to the visiting node as properties retrieved from source code. As discussed in the last section, there might be more types of properties that can reduce the number of nodes to be visited.

The result of further investigations on the kinds of useful properties of source code also affects the parsing way of source code itself. It will change the requirements of the source code parser. For example, some optional high cost feature of parsing will be executed by default. It will also change the database schema for the parse results to retrieve the useful properties with less cost.

6. RELATED WORKS

A bidirectional search algorithm in [6] or alike is called front-to-back algorithm. It uses the heuristic function to estimate the distance from the current visiting node (i.e., front) to the goal node (i.e., back). The function is similar to the heuristic function of A* search method [10]. The algorithm is executed under the assumption that the heuristic function estimates a value that is

equal to or less than the real value (i.e., not overestimating.) The function is said to be admissible if the property holds.

A bidirectional search algorithm called front-to-front [7] [8] uses the heuristic estimate function that calculates the distance from the current visiting node (i.e., front) to the frontier node of the opposite direction search (i.e., (another) front.) The algorithm achieves the best performance when the function is admissible and consistent, that is, given nodes x, y, and z where x and z are the end nodes of a path and y is on the path, the estimated distance between x and z is equal to or less than the summation of the real distance between x and y and the estimated distance between y and z.

The former algorithm has been verified by experimental evaluations [6] for at least Fifteen-puzzle problems. The latter has been proved theoretically in [8].

For call graphs, however, either type of heuristic function has not been found yet to the best of our knowledge. Hence bidirectional search algorithms with heuristic estimate functions cannot apply to call graphs.

Although unfortunately we have not found previous works on call graph traversal especially related to bidirectional search, researchers of call graph visualizer made a comment on bidirectional search [11]. From experiences of participants attending to a lab study to evaluate the visualizer, they said "A significant barrier to static traversal were event listeners, implemented using the Observer Pattern. To determine which methods were actually called, participants would have to determine which classes implemented the interface and then begin new traversals from these methods." Our approach could resolve the difficulty and might make their traversal processes easier.

7. CONCLUSION

We have offered a bidirectional search algorithms for a call graph of too huge source code to store all parse results of the code in memory. It reduces 8% of path extraction time for a case in which ordinary path search algorithms do not reduce the time. The algorithm refers to a method definition in source code corresponding to the visited node in the call graph. The significant feature of the algorithm is the referred information is used not in order to select a prioritized node to visit next but in order to select a node to postpone visiting. They contribute to the search time reduction.

ACKNOWLEDGEMENTS

We would like to thank all our colleagues for their help and the referees for their feedback.

REFERENCES

[1] Ryder, Barbara G., (1979) "Constructing the Call Graph of a Program", Software Engineering, IEEE Transactions on, vol. SE-5, no. 3, pp216–226.

[2] "Android Open Source Project", https://source.android.com/

[3] Scientific Tools, Inc., “Understand ™ Static Code Analysis Tool”, https://scitools.com/

[4] “Doxygen”, http://www.stack.nl/~dimitri/doxygen/

[5] Pohl, Ira, (1969) “Bi-directional and heuristic search in path problems”, Diss. Dept. of Computer Science, Stanford University.

[6] Auer, Andreas & Kaindl, Hermann, (2004) “A case study of revisiting best-first vs. depth-first search”, ECAI. Vol. 16.

[7] de Champeaux, Dennis & Sint, Lenie, (1977) “An improved bidirectional heuristic search algorithm”, Journal of the ACM 24 (2), pp177–191, doi:10.1145/322003.322004.

[8] de Champeaux, Dennis, (1983) “Bidirectional heuristic search again”, Journal of the ACM 30 (1), pp22–32, doi:10.1145/322358.322360.

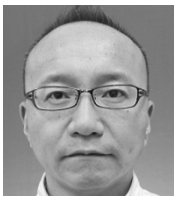
[9] Kwa, James B. H, (1989) “BS∗: An admissible bidirectional staged heuristic search algorithm”, Artificial Intelligence 38.1, pp95-109.

[10] Hart, Peter E. & Nilsson, Nils J. & Raphael, Bertram, (1968) "A Formal Basis for the Heuristic Determination of Minimum Cost Paths", IEEE Transactions on Systems Science and Cybernetics SSC4 4 (2), pp100–107, doi:10.1109/TSSC.1968.300136.

[11] LaToza, Thomas D. & Myers, Brad A, (2011) “Visualizing Call Graphs”, 2011 IEEE Symposium on Visual Languages and Human-Centric Computing, pp117-124, doi: 10.1109/VLHCC.2011.6070388.

AUTHORS

Koji Yamamoto works at Fujitsu Laboratories, Japan, since 2000. He is interested in software engineering especially using program analysis, formal methods, and (semi-) automatic theorem proving and its systems. He received doctoral degree in engineering from Tokyo Institute of Technology, Japan, in 2000. He is a member of ACM.



Taka Matsutsuka received his M.S. degree in Computer Science from Tokyo Institute of Technology. He works for Fujitsu Laboratories Ltd., and is engaged in R&D of OSS analysis. He is a visiting professor at Japan Advanced Institute of Science and Technology and a member of Information Processing Society of Japan.

TOWARDS A NEW APPROACH OF DATA DISSEMINATION IN VANETS NETWORKS

Ouafa Mahma[1] and Ahmed Korichi[2]

[1]Department of Computer Science and Information Technology, Kasdi Merbah University, Ouargla, Algeria
[1]mahma.wafa@univ-ouargla.dz, mahma.wafa@gmail.com
[2]Department of Computer Science and Information Technology, Kasdi Merbah University, Ouargla, Algeria
[2]ahmed.korichi@univ-ouargla.dz

ABSTRACT

In the 2000s, ad hoc networks was developed and highly used in dynamic environment, particularly for inter- vehicular communication (VANETs : Vehicular Ad hoc Networks).

Since that time, many researches and developments process was dedicated to VANET networks. This was motivated by the current vehicular industry trend that is leading to a new transport system generation based on the use of new communication technologies in order to provide many services to passengers, the fact that improves the driving and travel's experience.

These systems require traffic information sharing and dissemination the example as the case alert message emitting allowing the driver to minimize driving risks. Sharing such information between vehicles helps to anticipate potentially dangerous situations, as well as planning better routes during congestion situations.

In this context, we are trying in this paper to model and simulate VANET Networks in order to analyze and evaluate security information dissemination approaches and mechanisms used in this type of networks in several exchanges conditions.This in order to identify their limitations and suggest a new improved approach. This study was conducted as part of our research project entitled "Simulation & VANETs", where we justify and validate our approach using modeling and simulation techniques and tools used in this domain.

KEYWORDS

VANET, data dissemination, simulation, modeling, analysis and performance study.

1. INTRODUCTION

Today, road safety has become one of the biggest Challenges in the world, especially with the high dependence of people on vehicles and the growth of traffic problems (congestion and accidents ...). This reason push the researchers to Find ways for mitigate these problems. Among these ways : the intelligent transportation systems (ITS) developed new systems based on the

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 13–23, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60202

emergence of new communication technologies in the automotive industry in order to provide passengers more services to enhance the driving experience and travelers.

These systems recently known as Vehicular Ad hoc NETworks (VANET), systems allow communications between vehicles in order to exchange and share relevant information ,in the form of different categories of applications. A safety application is one of these applications that have aroused great interest among researchers.

In addition, the VANET systems often require knowledge of road conditions such as road status, accidents and congestion situations, and therefore the transmission of warnings messages allowing the driver of the drivers , this for minimize the dangers of driving. The exchange of such information and its sharing with distant vehicles can also help to anticipate potentially dangerous situations and to plan better routes during congestion situations.

The objective of this paper is to build a simulation model for the analysis of data dissemination protocols in vehicular ad hoc networks, in order to propose a new approach of traffic data dissemination which remedy insufficiencies of the approaches currently used. Where we interest in warning messages delivered in VANET networks. For achieve a better diffusion of these messages, consequently try to find the best parameters as time of dissemination and network overhead rates.

2.VANET NETWORK

Vehicular networks are a new class of wireless networks that have emerged by means of advances in wireless technology and the automotive industry. These networks also known with name of VANETs, which are considered as one of the real applications of ad hoc network, for communication between adjacent vehicles also between vehicles and stationary equipment.

The objective of VANET networks is to apply some notifications, such as dissemination of alert messages, reporting an accident between vehicles to reduce the probability of collision, the multimedia real-time applications and many other applications...

3. SHARING AND DISSEMINATION OF DATA IN VANET

Dissemination of traffic information is the principle of several research works, in consequence of the fact that the information is always shared concerns the traffic situation (for example the state of the road, condition of the car ...) in order to facilitate the movement of drivers and passengers on the roads, to enable them to take appropriate decisions to changes occurring in the road. In particular the information relevant to the risks and dangerous cases occur on the roads, this information necessities the immediate release and quick [1, 2, 3, 4] to ensure traffic safety.

This type of information is usually sent to a group of vehicles where the public interest here is consequently the most appropriate method is broadcast. although there are special cases where interest is concerned a specific group (eg: Geocast) (see Figure 1 and 2) .in general, the authors focus on methods Broadcast and Multicast, Geocast ... [1, 5, 6] according to their objective and their needs. In parallel with the fact that VANETs networks are mainly based on the ad hoc short-range communication between the vehicles to improve safety in vehicular environments (WAVE: iEEE 802.11p).

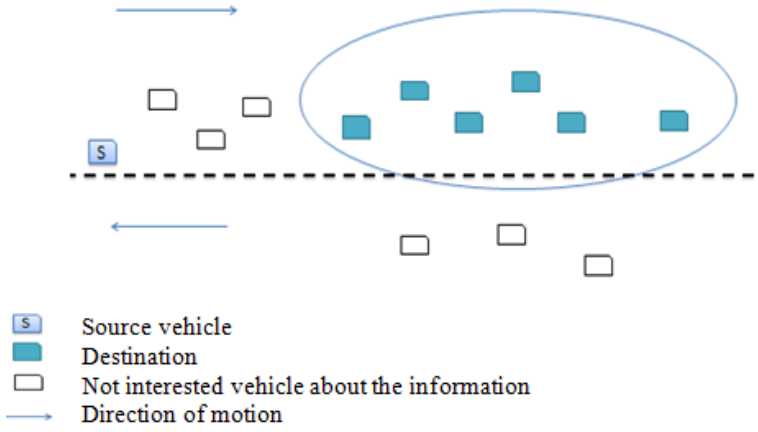


Figure 1. The information addressed to a specific group of vehicles

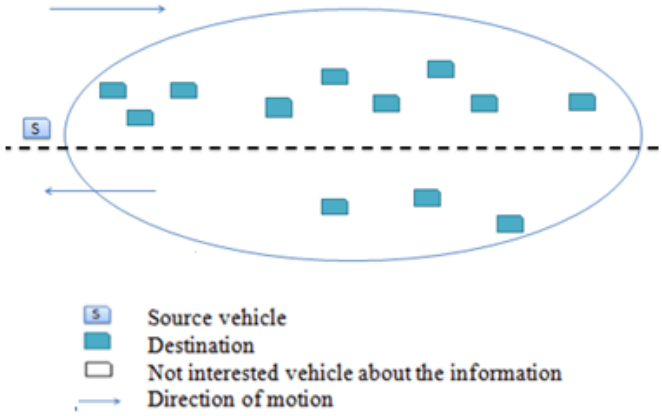


Figure 2. Information sent to all vehicles

Indeed, these previous studies use different concepts to achieve their goals and again to get the best results by a specific need (objective work). What is generally similar in these works is (1) the use of length of time. These shared information (warning messages, alerts messages ..) are issued only for a period of time (as long as the danger has existed), although there are very dangerous events cannot stand the wait and require the immediate release, this idea is not entirely absent, though his realism that is somewhat missed. (2) The lack of attention to the nature of the danger, for example:

• Loss of control over a truck carrying a flammable liquid, resulted in this slippage vehicle on Highway contains dozens of vehicles where the movement is very fast (vehicle speed can exceed 100 km / h), this shift is 100 m from the place of the accident (in Algeria).

• Another example: an accident occurs several times in European pays (like Belgium), the loss of control of the vehicle because of a breakdown on the level of the device named the limiting device and speed regulator where the vehicle was blocked at the speed of 200 km/h on the road.

• The results of these accidents were catastrophic, and that because of the lack of availability of the diffusion of this information in real time on the road.

Thus our objective in this work is not far from the mentioned points, we propose a communication system Geocast inter vehicles that enable the dissemination of safety messages (warning) in real time. Our objective is not the maintenance of the diffusion, but the reception and to send it message in time (in real time) potentially possible. This innovation based on the dissemination of the message of warning in the form of an alarm, in order to ensure a security service.

4. SIMILAR WORK

4.1. Urban Multi –hop Broadcast (UMB)

This protocol proposed in [7] designed to solve problems: Broadcast Strom, hidden nodes and which are related to the massive distribution in urban environments. This UMB protocol does not guarantee the absence of collusion, since it is possible to exist in more than one vehicle in the same segment. They can send in the same time the CTB. Again, this protocol works by some form of wait can be very long to select the next transmitter.

4.2. Smart Broadcast (SB)

It is a UMB improvement, wait time to assign the next re-broadcaster (relay) . with a delay function (WR) , the farther vehicles from the source always have a waiting period when this latter is finished the vehicle transmitted the packet CTB to the source. once the source successfully receives this packet, it then transmits the data packet. This SB protocol as it is Indicated, it is an improvement of UMB at latency term .a shorter waiting period that is in UMB.

4.3. Position-Based Adaptive Broadcast (PAB)

The authors of SB propose the PAB protocol (Position-Based Adaptive Broadcast) [8]which is based on the waiting time for relay vehicle before forwarding the information to improve the techniques proposed for the facility access and efficiency of broadcasts. Contrary to SB which works at the distance between transmitter and receiver, the PAB uses a formula calculated by the vehicle position and speed to find the delay time value.

4.4. Efficient Directional Broadcast (EDB)

This protocol [9] is based on the direction one using bidirectional antennas of the vehicles. It works in same principle UMB. To solve the problem of retransmission number redundant latency is proposed for each node in the sector the emission before the retransmission, this time is different for each vehicle, it is calculated according to the distance between the vehicle and the transmitter.

4.5. Reliable Method for Disseminating Safety Information (RMDSI)

The authors in [10] the authors selected the delay time to differentiate the priority of retransmission for each vehicle. It is to function similarly with the previous protocol, when the vehicle receives a packet, a waiting time before rebroadcast it. After the expiration of the waiting time, the vehicle retransmits the packet. Vehicles intending to duplicate the rebroadcast before their waiting time expires cancel their broadcasts. By simulation, when the network is highly

fragmented, RMDSI powerful that the UMB, which does not take the problem of the fragmentation of the network.

4.6. Reliable Broadcasting of Life Safety Messages (RBLSM)

In [11] the priority in the selection of relay is given to the vehicle with the vehicle nearest to the transmitter. The reason moreover nearer is more reliable. For example, a closer vehicle is supposed to have an intensity of better received signal. This protocol also uses the use of the packages of order and RTB CTB. The performances of the protocol are evaluated by simulation; However, only latency of one hop is provided. [3][11]

4.7. Multi-hop Vehicular Broadcast (MHVB)

This protocol [12] is similar to the precedents; it waiting time-based, such as the furthest vehicle from the transmitter always have the shortest waiting time. This time is calculated by an allocation function is not explicitly stated. After the expiration of the waiting time, the vehicle retransmits the packet. The protocol can be detecting traffic congestion. As each vehicle uses the number of its neighbors and its speed as an indication of congestion. [3]

4.8. Abiding Geocast (AG)

It is a system geocast of communication inter vehicle for the dissemination of massages of warning in VANET network. This model is proposed in order to ensure the dissemination of warning in order to: increase the probability of access to all relevant vehicle and reduce the overhead. This system uses different notions of time and space. To ensure inform group of vehicles exist in a geographical area someone on a risk proximity. Of which, the first vehicle detects the risk starts broadcasting a warning to other vehicles to inform them of this dangerous situation. In this work the authors used different dissemination strategies: (1) To achieve the first goal, they used the vehicle as reverse relay (2) Second goal: to update a time Waiting dynamically active vehicle for the next release. [1]

4.9. Optimal multi-hop broadcast protocol for vehicular safety (OCAST)

This protocol is an optimization of alert message dissemination in the VANET network for secure roads. Optimality in terms of time and number of transmission: using a dissemination strategy exploits the vehicle as opposed relay based on waiting times. And to complete the coordination of reliable and efficient distribution, smart periodic broadcasts to effectively adapt to VANETs networks. [5]

5. OBJECTIVES AND METHODOLOGY OF THE STUDY

Our mains objectives in these papers - as we indicated – are:

- Study of different dissemination protocols used in VANET networks .
- Create a model of simulation for applied the comparison and analysis performance of two studies in many cases: two scenarios and several parameters by analyzing the simulation results and found parameters.

- Provide an optimal solution for the dissemination of safety messages on the roads.

This is generally for measuring the flexibility of these systems accidents that may occur suddenly in roads whatever its nature. Because of the real measurements are not possible, we chose to use in this work the analytical study based on simulation results.

6. PERFORMANCE EVALUATION

This study is a result of comparison and analysis of two studies of performance in many cases two scenarios and several parameters by analyzing the simulation results and found parameters. An optimal solution inspired by these studies for time dissemination and diffusion rate.

Although there are other protocols in addition to previously indicated [14], [4] ...., we chose OCAST and AG since they use parameters give potentially effective results compared to other proposes .En even more the OCAST is already compared with other in [2].

A. Metrics

Parameters measured are the rate of informed vehicles before the risk area, and the probability of access to all vehicles concerned in two different cases of risk: (risk is located in a random manner (case 2) and others in a predictable manner for the system (Case 1)). for a short duration time and long duration of time.

By simulation tools exists, our model is not much different from the used in [5] with a simple change in parameters for measure the effectiveness of the two systems in case of changing of risk on the road. The found results can be quite sufficient to compare and study the performance of these two protocols. .Generally OCAST is an improvement of what studied.

Indeed, we have chosen to add these results broadcasts capabilities used in [5] to the Geocast capacity proposed in [1] to provide an improved system for the dissemination of traffic information. So our model is defined in the next section but realistic simulation results of our approach are defined in details in our next production with a new algorithm for the message diffusion process in VANETs, where we use the simulator OMNET ++ (Version: 4.3.1) with the traffic simulator SUMO (0.17.1) and VEINS (2.2) to achieve our results.

B. Simulation

In this party, we use OMNET ++ open source simulation library that is written in C ++. This discrete event simulator simulate both types of networks (wired and wireless), in which the different network nodes can communicate via messages. The OMNET ++ tool has several advantages over others: simple to use, easy to learn by means of its user interface and its generic structure. It allows to find and obtain clear results, detailed and in many forms, diagram and drawings. This simulator is flexible to make changes in the created simulation models and even reuse these models. And more recently, this simulator is widely used in VANET network simulation domain, in practically protocols simulation model related to the MAC layer. Consequently, we prefer to use this simulation tool to avoid several problems related to the implementation of the simulation model by other simulators, especially when a large number of messages exchanged and shared between network nodes.

B.1. Simulators information's

Table 1. Informations about the Simulators.

Simulator	Version
OMNET++	4.1.1
Sumo	0.17.1
Veins	2.2

B.2. The simulation parameters

To achieve the simulation scenarios, we chose the following parameters:

Table 2. Simulation parameters.

Description	value
Transmission range(R)	250m
Straight road	7 km
Mac layer	IEEE802.11
Safety distance	250m
Effect distance	10Km
traffic volume λ	200 ~ 700 (veh/hr)
Speed mean(Smean) Speed variation ε	30m/s 5m/s
location of the warning	Case 1 = 0meters Case 2= variable
simulation time	3000s
Start time of the warning event	Case 1 = 400s Case 2= variable

As we say, we chose the same simulation parameters with slight changes, this for confirmation of the results. Except the second case where we tried to stay away from the private of fixed location of event for this risk to measure the effectiveness of the two systems. We omit the start time of the warning event to let the system reach a stable state considering the distribution of vehicles over the road. For both scenarios, when the warning event occurs, the beginner of dissemination is at the location of the safety line.

C. Results and discussion

Case 1:

In figure 4: different traffic value (measured density 1 to 15 vehicles are informed to 100% before the risk area (both systems). But imperfection AG appeared in (Figure 3) before arriving at risk (after the security line) this rate is a little different (the existence of vehicles after they inform exceed the safety line): a rate (> 20%) and (<40%) of vehicle are not inform in the low density network. But for OCast delivery rate of 100% for all densities.

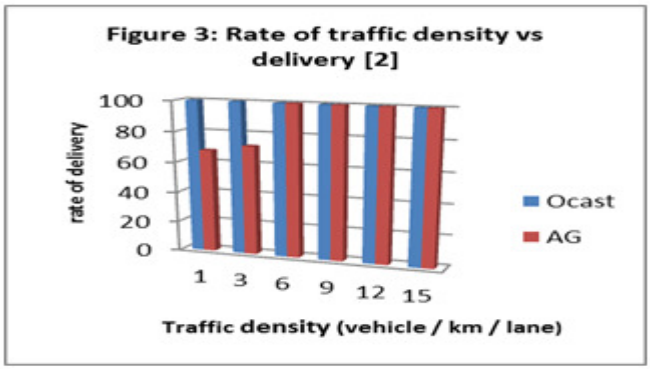


Figure 3. Rate of traffic density vs. delivery [5]

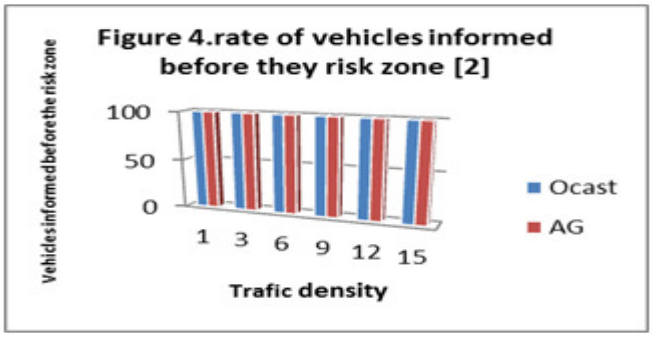


Figure 4. Rate of vehicles informed before risk zone [5].

Case 2:

In this case the simulation results are slightly different. In FIG 5: different traffic value (higher density then 7 vehicles are informed to 100% before the risk area (both systems).

But imperfection of both AG and Ocast system appeared in before arriving at risk (after the security line) and the beginning of the appearance of risk: the vehicles after they inform exceed the safety line, although this rate is not large but it is important in the warning information dissemination process.

Generally for all densities (Figure 3), OCast shows better performance than the AG as sending message "Stop" guarantee control the number of alerts issued in the network consequently it ensures minimization overload network.

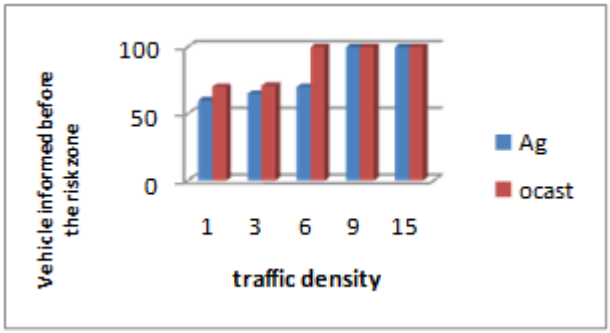


Figure 5.Rate of the vehicles before they risk area

But the case of an unforeseeable risk, we see a better performance for both protocols in a density more than 6, but this performance is decreased to the density less than 6 at the beginning of the appearance of risk.

We note that in the determination of the nature of risk and the results were good. Unlike in case the unknown risk (the location and characteristics) make traffic information sent uncertain and imprecise. Indeed, the simulation results show the effectiveness of OCAST compared to AG (case1) and but these results do not indicate the futility and uselessness of AG or OCast (case2).

Generally the purpose of our study is to measure the flexibility of these systems accidents that may occur suddenly in roads whatever its nature.

D. Our proposed approach

Our proposal is defined as the GODD "Geocast Optimization for Data Dissemination in vanet networks", a Geocast protocol introduced as an improvement time (the emission time) and number of transmissions (overhead), a flexible system to change the road for a better dissemination of safety data in VANET networks.

We try in this work as much as possible to ensure the immediate and optimal delivery of the warning information to all vehicles that are near to the risk, consider the validity of the message (the life time) and the direction of the mobility of vehicles in the network.

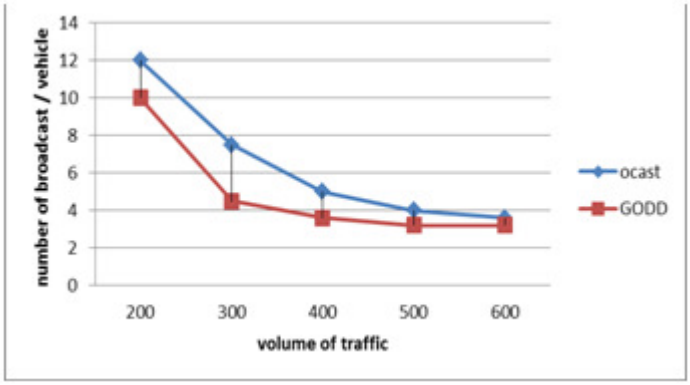


Figure 6. Rate of the vehicles before they risk area

This communication system Geocast inter vehicles is interested in the warning information released as soon as possible quickly the source node to several recipients in a region geographically define to secure these vehicles.

We use strategies to achieve our objective, to inform all these points while reducing unnecessary broadcasts and ensure a favorable reception time.

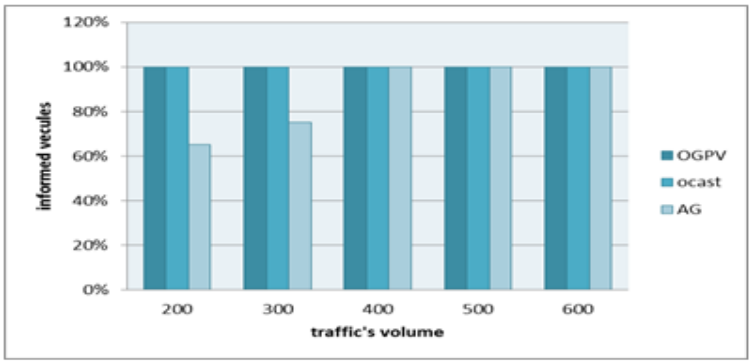
Broadcasts are organized in this system by the following process:

1- Initialization: when a risk occurs, the damaged vehicle will broadcast a warning message to neighboring vehicles (in its transmission area).

2- Selection of vehicle drivers:

The relay vehicle is an active vehicle chosen by its proximity to the transmitter and its relative position in the latter.

3- Exit the system: The made to select a new relay, the source must receive a message of "stop" this relay to indicate you can leave the area.



7. CONCLUSIONS

We tried to present in these papers an inter vehicle communication system Geocast: GODD as an optimization of the proposed approaches. Which we introduce a new contribution to more secure passengers in the roads.

This work is a description of what our approach detailed in our next production, where we show the optimality of the system by the simulation tool by OMNET ++ simulator used (version 4.3.1) and SUMO (Version 0.17) with VEINS.

The system offer better message delivery performance side of caution, the thing that evolve road services for VANETs.

REFERENCES

[1] Y.Qiangyuan Yu and H. Geert (2008) “Abiding Geocast for Warning Message Dissemination in Vehicular Ad Hoc Networks”, IEEE International Conference on Communications Workshops (ICC) Beijing, China, pp. 400-404.

[2] M. Koubek, S. Rea, D. Pesch (2008) “Effective Emergency Messaging in WAVE based VANETs” Centre for Adaptive Wireless Systems,Electronic Engineering Dept, Cork Institute of Technology, , in Proc. First International Conference on Wireless Access in Vehicular Environments (WAVE 2008), Dearborn, MI, USA.

[3] S. Panichpapiboon, and W. Pattara-atikom (2011)“A Review of Information Dissemination Protocols for Vehicular Ad Hoc Networks”, IEEE.

[4] N. Haddadou (2014) " Réseaux ad hoc véhiculaires : vers une dissémination de données efficace, coopérative et fiable " (doctorate These).

[5] A. Benaidja ,S. Moussaoui and F. Naït-Abdesselam (2013) “An Optimal Broadcast of Warning Messages in Vehicular Ad Hoc Networks “, International Journal of Computer and Information Technology , ISSN: 2279 – 0764, Volume 02– Issue 05.

[6] C. Maihofer, T. Leinmuller, and E. Schoch (2005) “Abiding geocast: time–stable geocast for ad hoc networks”, in Proceedings of the 2nd ACM international workshop on Vehicular ad hoc networks (VANET’05), New York, USA, pp. 20-29.

[7] G. Korkmaz, E. Ekici, F. Ozguner, and U. Ozguner (2004) Urban multi-hop broadcast protocol for inter-vehicle communication systems. In ACM International Workshop on Vehicular Ad Hoc Networks, New York, NY, USA.

[8] Y.T. Yang and L.D. Chou (2008) "Position-Based Adaptive Broadcast for Inter-Vehicle Communications," Communications Workshops, ICC Workshops '08. IEEE International Conference on , vol., no., pp.410-414, 19-23.

[9] D. Li, H. Huang, X. Li, M. Li, and F. Tang (2007) “A distance-based directional broadcast protocol for urban vehicular ad hoc network,” in Proc. IEEE Int’l Conf. on Wireless Comm., Networking and Mobile Computing (WiCom), Shanghai, China, pp. 1520–1523.

[10] S. Khakbaz and M. Fathy (2008) “A reliable method for disseminating safety information in vehicular ad hoc networks considering fragmentation problem,” in Proc. IEEE Int’l Conf. on Wireless and Mobile Communications (ICWMC), Athens, Greece, pp. 25–30.

[11] M. Taha & Y. Hasan (2007)“VANET-DSRC protocol for reliable broadcasting of life safety messages,” in Proc. IEEE Int’l Symp. on Signal Processing and Information Technology, pp. 104–109.

[12] T. Osafune, L. Lin, & M. Lenardi (2006) “Multi-hop vehicular broadcast (MHVB),” in Proc. IEEE Int’l Conf. on ITS Telecomm. (ITST), Chengdu, China, pp. 757–760.

[13] Q. Xu, T. Mak, J. Ko and R. Sengupta (2004) “Vehicle-to vehicle safety messaging in DSRC”, in Proceedings of the First ACM workshop on Vehicular ad hoc Networks (VANET).

[14] Harshvardhan P. Joshi (2006) “ Distributed Robust Geocast:A Multicast Protocol for Inter-Vehicle Communication”, Computer Networking - Electrical Engineering, Raleigh these.

AUTHORS

Ouafa MAHMA

Obtained on testimonies license (2006) and master (2011) in the fundamental automatic information from the University of Ouargla in Algeria .Prepare for a doctoral degree in the field networks and informatics systems in the same university.

Ahmed KORICHI

Obtained on testimonies doctorate (2004) in automatic information from the University of Batna in Algeria. Prepare for a professor degree in university of Ouargla in Algeria .

INTENTIONAL BLANK

MODELLING DYNAMIC PATTERNS USING MOBILE DATA

Suhad Faisal Behadili[1], Cyrille Bertelle[1] and Loay E. George[2]

[1]Normandie Univ, LITIS, FR CNRS 3638, ISCN, ULH, Le Havre, France
suhad.behadili@etu.univ-lehavre.fr
cyrille.bertelle@univ-lehavre.fr
[2]Baghdad University, Computer Science Department, Baghdad, Iraq
loayedwar57@scbaghdad.edu.iq

ABSTRACT

Understanding, modeling and simulating human mobility among urban regions is very challengeable effort. It is very important in rescue situations for many kinds of events, either in the indoor events like evacuation of buildings or outdoor ones like public assemblies, community evacuation, in exigency situations there are several incidents could be happened, the overcrowding causes injuries and death cases, which are emerged during emergency situations, as well as it serves urban planning and smart cities. The aim of this study is to explore the characteristics of human mobility patterns, and model them mathematically depending on inter-event time and traveled distances (displacements) parameters by using CDRs (Call Detailed Records) during Armada festival in France. However, the results of the numerical simulation endorse the other studies findings in that the most of real systems patterns are almost follows an exponential distribution. In the future the mobility patterns could be classified according (work or off) days, and the radius of gyration could be considered as effective parameter in modelling human mobility.

KEYWORDS

Modelling, Armada, Probability Distribution, Inter-event time, Displacements

1. INTRODUCTION

The purpose of simulation analysis is to acquire and analyze the results in well conceptual vision, in order to give high indications for decision makers, pivoting on two events kinds, the discrete events and frequent (continuous) events. In order to explore the mobility characteristics, eventual real effects conditions, and actions of a specified system, they should be modeled mathematically [1, 2, 3, 4, 5]. However, the most important and difficult issue in modeling and simulating any system is the determination of the probability distribution and parameters to model the uncertainty of the system input variables. Many researches used CDRs to study the collective and individual human mobility (Gonzalez et al., 2008), the segmentation of urban spaces (Reades et al. 2009), understand of social events (Calabrese et al., 2010). (Phithakkitnukoon et al., 2010) suggested the activity-aware map, using the user mobile to uncover the dynamic of inhabitants, for urban planning, and transportation purposes. (Ratti et al., 2010) deal with large telecommunication database for Great Britain in order to explore human interactions, and emphasized on the highly interactions correlation with administrative regions [10].

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 25–30, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60203

2. DATA SET

The case study data is Call detail Records of mobile phone, composed of 51,958,652 CDRs, represent entry records of 615,712 subscribers, for the period starting from 4 (Friday)-15 (Tuesday) of July in 2008. However, it contains individuals occurrence in discrete (irrelevant) mode only, means that any mobile individual activity is recorded at (start/end) time, but there is a lack of information, which is supposed to indicate the individual's occurrence during inactive case (mobility without any mobile phone activity). There is no data meanwhile the mobile phone is idle, i.e. inactive or doesn't make any communication activities neither calls nor SMS activities. As well as, the available spatial data is only of the towers (X,Y) coordinates, hence it would be considered to estimate individual's transitions from position to another (from tower to tower), hence the positions would be determined approximately with regarding to the tower coverage (signal strength). With regarding to the non-deterministic & discrete nature of this data, therefore the collective behavior would be the effective approach to be analyzed and simulated. Since, each individual could be disappeared for a while from the DB, which makes individual tracing is unworthy, without significant indications on the people behavior in the city [3].

3. DESCRIBING INDIVIDUALS' ACTIVITIES

The communications activities have heterogeneous nature, since the individuals are varied in their usage of mobile phone, which are ranged between (rarely-frequently) usages during specific period. The individuals are grouped according to their total activities. However, probability of waiting time (inter-event time) ΔT of each two consecutive activities has been computed for each individual, so the individuals would be grouped with regarding to their activities. Computing the probability function is to get the system universal behavior (pattern), according to consecutive inter-event times, where most life systems are modeled by exponential law [1, 6, 7, 5, 8]. So, the distribution of the average inert-event time ΔT_a is estimated by exponential distribution law as in equation (1), its histogram in figure 1. The demonstration in figure 2, reveals that the longer waiting times are characterizes the individuals of fewer activities.

$$P(T) = (\Delta T) exp^{-\Delta T} \qquad (1)$$

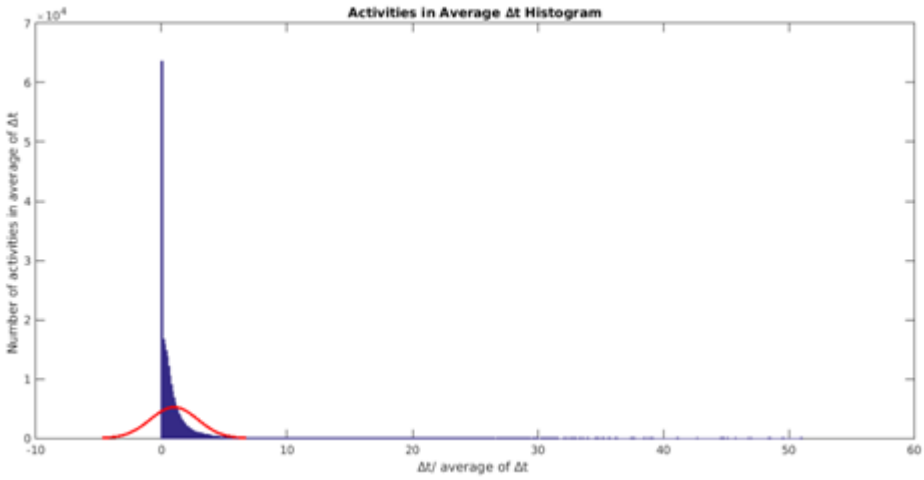


Figure 1: Histogram of Activities in Average of Inter-Event Time

4. INTER-EVENT TIME OBSERVATIONS

The modeling and simulation process starts by computing probability model, which is modeled the data and simulate the responses of the model, using one of the well-known functions, the probability density function PDF [3, 4, 9]. The main parameter in this study is the mean of inter-event time, where the means of samples of the distribution are drawn as exponential distribution, then compute the mean of all means. In order to understand the behavior of sample means from the exponential distribution, to get the universal system pattern (general population law). The computations are done as follows:

1. Manipulate all 12 days data, each day independently due to daily regular patterns, for all users (individuals) in spatio-temporal manner. Then eliminate the individuals of only one occurrence in the CDRs, since they didn't have significant indication on mobility.

2. Sorting the data by time, in order to have the real sequence of positions transitions of individuals' trajectories. Classify the data according to individuals' activities (sampling), by computing inter-event time Δt (waiting time), where it is the time elapsed between each successive activities of each individual, which is ranged between 15-1440 minutes, this sampling is done according to logical intuition, since 15 minutes is the minimum time that can give mobility indication, and the 1440 minutes (24 hours) could be considered as the highest elapsed time to travel inside the observed region.

3. Compute the inter-event time of all individuals ΔT, then compute ΔT_a (average inter- event time) of all individuals, then classify (min, max) samples according to activities score (activities densities), then compute$P(\Delta T)$. However, $\Delta T/\Delta T_a$ is the average inter-event time of the all individuals (whole population).

4. Compute exponential distribution probability for each day, then for the whole days as in figure 2 to identify the universal population pattern law, then to show the approximation of all 12 days curves with the curve of their average values as in figure 3.

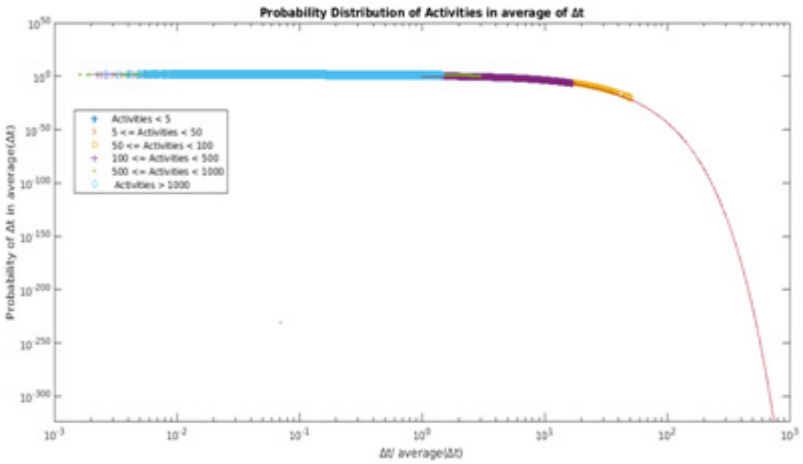


Figure 2: Waiting time distribution P(Δt) of mobile activities, where Δt is spent time between each two successive activities, legend symbols are used to distinguish the individuals' groups according to their activities ratio, for whole population during whole period, activities in average Δt.

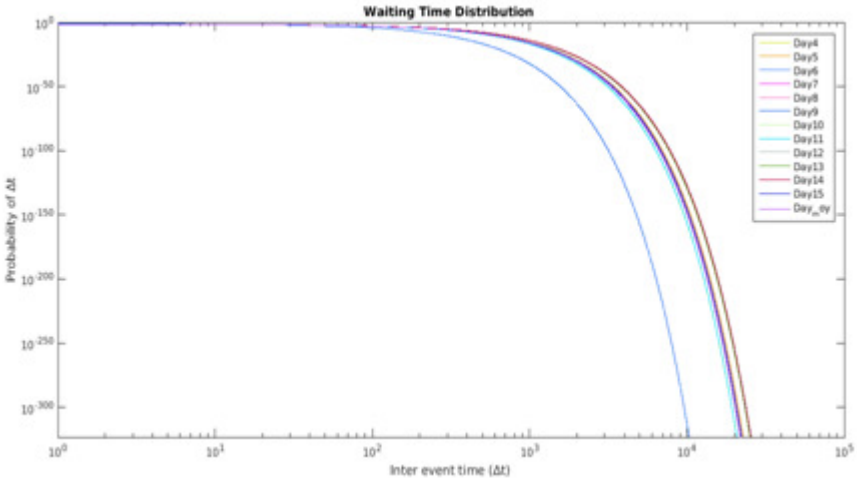


Figure 3: Waiting time for the 12 days curves with the curve of their average.

Exponential distribution (probability distribution) in equation (1) is capable of modeling the events happened randomly over time. In this case it describes the inter-event time, and the average of inter-event time of individuals' activities [1, 2, 3, 9]. Hence, the cutoff distribution is determined by the maximum observed inter-event time (significant parameters), which the individual can wait to make any mobile activities, where it is Δt=1431 min. The resulted law as in equation (1) is computed according to algorithm of complexity $O(n^3 + 2n^2 + n)$. The distribution shows that short inter-event times have higher probability than long ones, and the 12 days have similar patterns of activities. As well as, by following the same manner mentioned above, in order to compute the displacement statistics of all the individuals during the whole observed period, the displacement probability distribution is computed. However, the Δr is the traveled distance between each two successive activities during time ΔT_0 of the range 20-1440 minutes, and the P(Δr) is the displacement distribution, the investigated distances would be limited by the maximum distance that may be traveled by individuals in the ΔT_0 (time intervals). Hence, the cutoff distribution is determined by the maximum observed distance, which individual can travel is Δr=7.229515 Km along the day hours, since the maximum time slice couldn't exceed 24 hours with regards to observed region. The displacements distribution is approximated by power low as in equation (2). The resulted distribution shows as in figure 4 that the displacements are clear in 10^4 m, whereas decreased after this value, the Δr is ranged between (0-1) Km.

$$P(\Delta r) = (\Delta r) exp^{-(\Delta r)} \qquad (2)$$

Well, the traveled distance (displacements) distributions of the 12 days curves with the curve of their average are demonstrated in figure 5, they have almost identical patterns. Hence, the obtained averages of the waiting time distribution are max Δ T = 1431 minute and min ΔT =0 minute, whereas the max Δr =7.229515e+04 meter and min Δr =0 meter.

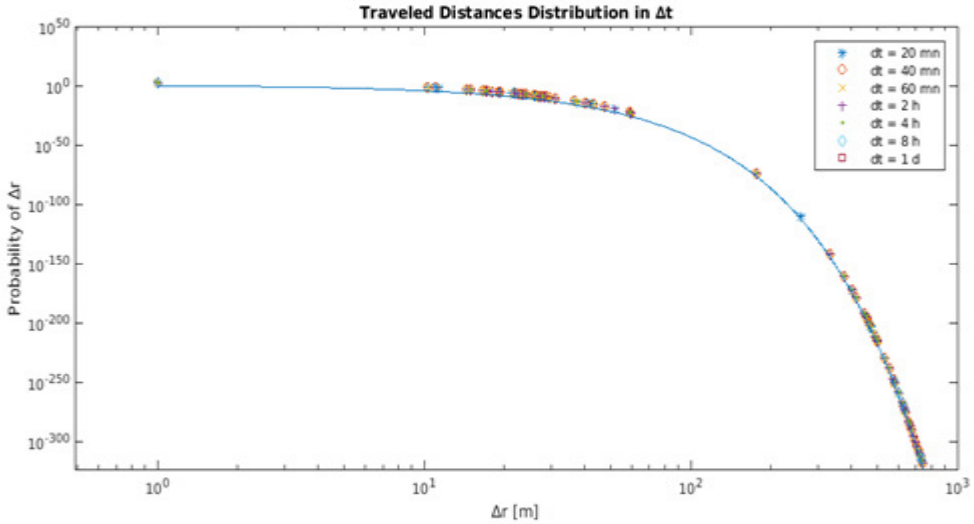


Figure 4: Distances probability distribution (displacements) P(Δr) for waiting times (inter-event times) Δt_0 for one day, cutoff distribution is determined by the maximum distance traveled by individuals for specific Δt_s.

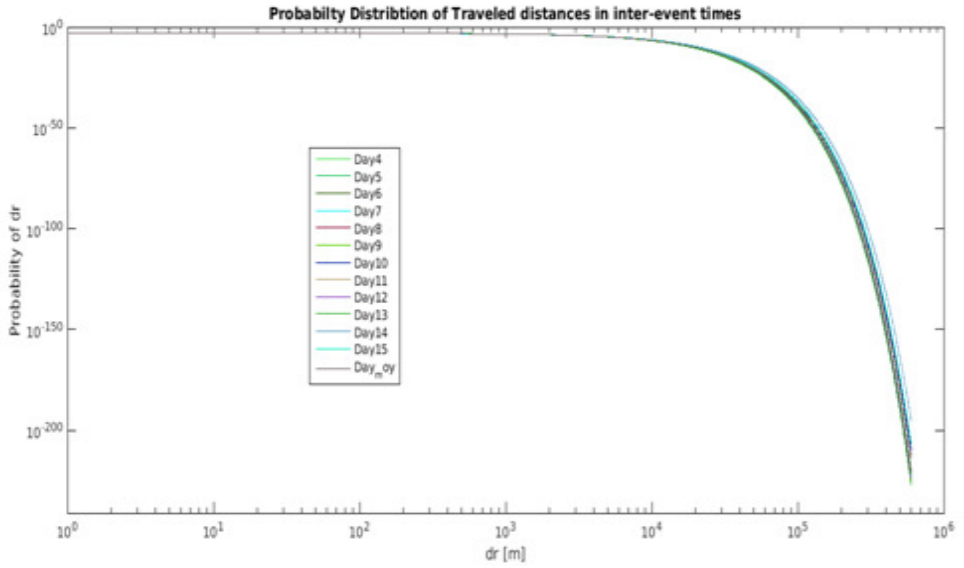


Figure 5: Displacements for the 12 days curves with the curve of their average.

5. CONCLUSIONS

The great attending recently towards dealing with catastrophic disasters, cities planning, diseases spreading, traffic forecasting, decision making for rescue people with disabilities in emergencies ...etc. [10], all these activities need working on huge data and a geographic data, to describe human behavior and mobility patterns, for giving the simulation model more reality, since almost simulation platforms are deal with only a grid to be the background of the model, so the resulted model will be far away from reality.

The results of this investigation show that the inter-event time Δt between each two successive activities has bursty pattern, since there is long period without activities, this gives indication about the population's heterogeneity. And human trajectories follows power-law distribution in step size Δr, and they are modeled using displacements and waiting time distributions, as well as the mobility traveling patterns are show that there are many short distances in contrast little long distances. The mobile phone data is used to reconstruct the population trajectories. However, the patterns of all days are very similar and they have approximately identical curves either in spatial or temporal features. As next forward step with this kind of studies the mobility patterns would be classified with regarding to the (work or off) days for more understandings to the life patterns. Whereas, the radius of gyration would be considered as significant modelling parameter to give the model more reality with focusing on patterns regularity.

REFERENCES

[1] CS302 Lecture notes, (2015) "Simulating Random Event", http://web.eecs.utk.edu/~leparker/Courses/CS302-Fall06/Notes/PQueues/random_num_gen.html

[2] "Simulation: An Introduction", (2015) Stat. Chapter 10, http://www.stat.nus.edu.sg/~zhangjt/teaching/ST2137/lecture/lec%2010.pdf

[3] Sougata Biswas, (2014) "Simulation of Exponential Distribution using R",https://rstudio-pubs-static.s3.amazonaws.com/26693_e1151035722942b2813c0063c6b220ae.html

[4] Vincent Zoonekynd, (2015) "Probability Distributions", http://zoonek2.free.fr/UNIX/48_R/07.html

[5] Solver, (2015) "Simulation Tutorial Introduction", http://www.solver.com/simulation-tutorial

[6] Siqi Shen, Niels Brouwers & Alexandru Iosup, (2011) "Human Mobility in Virtual and Real Worlds: Characterization, Modeling, and Implications", report number PDS-2011-007, ISSN 1387-2109.

[7] The Pennsylvania State University, (2015) "Probability Theory and Mathematical Statistics", STAT 414/415, https://onlinecourses.science.psu.edu/stat414/node/97.

[8] Ilya Narsky, Frank C. Porter, (2013) "Statistical Analysis Techniques in Particle Physics, Fits,Density Estimation and Supervised Learning", Wiley-VCH, 1 edition.

[9] Marta C. Gonzlez, Cesar A. Hidalgo & Albert-Laszlo Barabasi, (2009) "Understanding individual Human Mobility Patterns", nature 453.

[10] Gunther SAGL, Bernd RESCH, Bartosz HAWELKA, Euro BEINAT, (2012) "From Social Sensor Data to Collective Human Behaviour Patterns-Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments", GI_Forum 2012: Geovizualisation, Society and Learning. Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-521-8.

PERSONAL IDENTITY MATCHING

Mazin Al-Shuaili[1] and Marco Carvalho[2]

[1,2]Florida Institute of Technology, Melbourne, USA
Malshuaili1994@my.fit.edu, mcarvalho@cs.fit.edu

ABSTRACT

Despite all existing methods to identify a person, such as fingerprint, iris, and facial recognition, the personal name is still one of the most common ways to identify an individual. In this paper, we propose two novel algorithms: The first one uses sound techniques to create a multi-dimensional vector representation of names to compute a degree of similarity. This algorithm compares names that are written in different languages (cross-language) and transliterated names in English. The second algorithm builds on the first one and measures the similarities between full names, taking into account the full name structure. We evaluate the algorithms for transliterated names and across languages using Arabic and English as an example. Significant results are achieved by both algorithms compared with other existing algorithms.

KEYWORDS

Border Security, Cross-Language Information Retrieval, Information Retrieval, Security Screening System.

1. INTRODUCTION

There are many methods to identify a person, including biometrics, such as fingerprint, iris, and facial recognition. Despite the availability of these techniques, the name of the person is more commonly used to identify an individual, especially in border control measures, criminal investigations, and intelligence analysis. The personal name is used when people apply for visas or on their arrival at a port (land-port, seaport, or airport). The International Civil Aviation Organization (ICAO) recommends all passports to contain personal names in the English language; therefore, countries have worked to provide the names of passport holders in the English alphabet. This transliteration causes a problem with spelling variations, such as with name "Mohammed," as shown in Table 1. In this case, some systems, such as security screening systems, need to deal with those differences when one-to-one comparison does not work correctly.

There are many screening systems that identify a given name, such as that of a traveler or suspect, with a list of names. The No- Fly List, for example, was created by the Transportation Security Administration (TSA) and is now used by the Terrorist Screening Center (TSC) in the United States [1]. This list contains the names of all people who are not permitted to travel to or from the United States. Also, the manifest is a document listing passengers and crew information that used to be sent by fax. Today, this information is sent electronically and managed by an

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 31–43, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60204

advanced passenger information (API) system. API sends the passengers' information (manifest) in real time to their destinations while they are checking in for their flights. A response comes from the destination with an indicator saying whether the passenger is allowed to board. In addition, Interpol is an intergovernmental organization that maintains a list of suspects that is shared among most countries. The names in the mentioned lists are checked with the private lists in each country or even each security sector separately using screening systems. Those systems might return a result of comparing two names as negatively matched (False Positive) or negatively unmatched (False Negative) as a consequence of name variations.

Table 1. Spelling Variations of Mohammed

Name	Nationality
MAHMMED	EGYPT
MOHMED	INDIA
MHOHAMMED	BANGLADESH
MUHAMMADE	INDIA
MAHOMED	SULTANATE OF OMAN
MUHMMET	TURKEY
MUHAMMET	TURKEY
MOHD	MALAYSIA

There have been many cases in which a person was arrested or stopped on suspicion of his or her name on the wanted lists. Senator Edward Kennedy was stopped serval times while boarding an aircraft because his name matched someone on the No-Fly List [2]. In another case, Congressman John Lewis was stopped for an additional security check. Sixty-tow-year-old Sister McPhee is an education advocate for the Catholic Church who was stopped repeatedly for nine months because her name matched that of an Afghani man whose name was included on the No-Fly List. There have been several occasions when infants and toddlers were stopped from boarding airplanes because their names negatively matched with names on the No-Fly List [2]. Screening systems must minimize these types of errors.

The side effect of screening systems is that common names are often written with different spellings for several reasons, including natural spelling variations, phonetic variations, and compound names [3], [4]. Transliteration causes spelling variations for a single name due to phonetic similarity [5], [6]. The names on the Interpol list and the API list are Romanized (transliterated). In 1918, Soundex was published as the first algorithm that tried to solve this type of problem by indexing names based on their sounds. Soundex studied the letters of the English alphabet and grouped them into seven groups, one of which is ignored by the Soundex algorithm. Section 2 discusses Soundex in more detail.

Cross-Language Information Retrieval (CLIR) also struggles to deal with proper names because they are out of vocabulary (OOV) [1], [7], [8], [9]. Furthermore, the authors in [7] demonstrated that about 50% of OOV words are proper names. Proper names might affect the performance of English-Arabic cross-language transliteration by 50% [10]. One solution for this problem is to transliterate all names from one language to another language so that all names have the same format (characters) to be compared, as applied in [10]. Transliteration of names from one language to another is not easy and is expensive due to phonetic variations in some languages [1], [6]. Also, a static solution, such as a dictionary, gives high performance but needs to be updated periodically with new names. When people borrow names from other cultures or invent new names, the number of names increases. We need a solution that can deal with these permanent

changes. Names in all languages must be standardized in order to be compared accurately and efficiently for CLIR.

The proposed algorithm, Personal Identity Matching (PIM), defines names as points in seven-dimensional space. The dimensions represent six groups of characters that are classified by Soundex plus a seventh group that represents the characters that are ignored by Soundex. Once PIM sets the names in space, then the distance is calculated between those points. The distance is converted to the similarity between 1 for an exact match and close to zero for no match. It is clear that PIM is unlike Soundex in how it compares names; PIM uses distance, whereas Soundex uses code. Our result shows that PIM has 33% better accuracy than Soundex and 11% better accuracy than all of the other algorithms involved our testing. PIM computes the similarity between individual names, but for a full name we create another algorithm (PIM-Full name or PIMF). PIMF breaks a full names into tokens (individual names), then it sends a pair of names into PIM, taking into account the full name structure.

2. RELATED WORK

Name matching is divided into two approaches. First, a phonic algorithm tries to avoid common problems of name variation by encoding names based on their sound [5]. Many phonetic algorithms have been introduced since Soundex, such as Metaphone, Cologne, NYSIIS**,** and Caverphone. They have gone through several stages since the early 1900s, when Margaret K. Odell and Robert C. Russell introduced the Soundex algorithm. The latest phonetic algorithm is Metaphone 3, created by Lawrence Philips in 2009 [11]. The same author created three versions of Metaphone, the third of which (Metaphone 3) was made for commercial purposes. Although all of the algorithms have different rules for converting a name into a code, they have common criteria: 1) They target a specific language or even for a specific database, such as Caverphone, which was created to match the elections lists generated between the late 19th century and early 20th century. 2) They generate code for each name: two names are matched only when both of their codes are the same; otherwise they are unmatched. 3) Most of them have codes of a fixed length, except Refined-Soundex has an unlimited length of code. Soundex has the fewest rules of all phonetic algorithms.

Soundex primarily groups characters based on their sounds, as shown in the next table (Table 2). The sounds of those letters are based on English pronunciation, and all of the letters' sounds are clustered into six groups. Other characters (Group 0), including {a, e, i, o, u, w, h, y}, are ignored unless they appear in the first character of the name. The Soundex algorithm keeps the first character without changing it because the first character is clearly pronounced; therefore, there are fewer errors made at the beginning of a name [12]. The Soundex code is a combination of a letter and three digits. The name "Jackson," for example, is converted to the code "J250."

In phonetic algorithms that generate code for each name, the comparison between names is a one-to-one relation (match or mismatch), and there is no concept of distance or similarity between two names. Therefore, alternative algorithms are needed to provide a better result that gives a degree of similarity between one string (*S1*) and the other string (*S2*). The most common algorithms that calculate a distance or a similarity are the Levenstein distance (edit distance), Jaro-Winkler, Smith-Waterman, Damerau-Levenstein, and N-gram algorithms.

Table 2. Soundex for English and Arabic

No	English Letters	Arabic Letters
0	A, E, I, O, U, W, H, Y	ا,ع,ي,هـ,ح,و
1	B, F, P, V	ف,ب
2	C, G, J, K, Q, S, X, Z	س,ش,ك,غ,ج,خ,ق,ص,ز
3	D, T	د,ض,ظ,ت,ذ,ث,ط
4	L	ل
5	M, N	م,ن
6	R	ر

Edit distance (Levenstein distance) is the minimum number of edit operations needed to convert S_1 into S_2; where each insertion, deletion, or substitution counts as one operation [13]. The minimum number of edits is divided by the longest string to calculate the distance between two strings. Damerau-Levenstein **distance includes** extra operation, a transposition, in addition to the other operations that are used by edit distance. The Smith Waterman algorithm is similar to the edit distance one but uses a phonetic algorithm to enhance the edit distance; e.g., if $a_i \approx b_j$, then the score is 2 (a_i&b_j are in the same group in Soundex table), or If $a_i = b_j$, then the score is 5; otherwise it is -5 [3], [14]. Jaro distance is based on common characters between two strings, S_1 and S_2, with half of the longest string used for transposition [3], [12]. It is commonly used with duplicate detection (data linkage) when a name needs to match other names within some distance threshold [3]. The Jaro-Winkler algorithm improves the Jaro algorithm by boosting the error score in the prefix characters. Some other algorithms that calculate distance divide a name into tokens such as n-gram and longest common sub-string (LCS). All those algorithms use a threshold to determine matching names.

The algorithms mentioned cannot deal with a full name structure. The full name is composed of more than one name usually a combination of a first name and a surname. In addition, a middle name (often the father's name) might be added between the first name and surname to increase the determination of personal identity. The middle name can be a father's name that is followed by his father's name, and more ancestors' names can be added recursively. In addition, in some countries the last name is followed by the first name, then the other names. We generalize the full name structure using a regular expression as follows: $F = n(m_i)^* l \mid (l, |l) n(m_i)^*$ where n is the first name, l is the last name, and m_i represents all other names in the sequence, starting from i=1; 1 is for the father, 2 is for the grandfather, and so on.

$$sim_{MongeElkon}(A,B) = \frac{1}{|A|}\sum_{i=1}^{|A|} \max\{sim'(a_i,b_i)\}_{j=1}^{|B|} \quad (1)$$

There are a few algorithms that deal with full names, and the most interesting one is the Monger-Elkan algorithm. Monge-Elkan distance measures the similarity between two full names that contain at least given names and last names. Each token (name) is calculated using a similarity function, which is $sim'(a_i, b_i)$, such as edit distance or the Smith Waterman algorithm. Then, the average number of tokens in A is computed. The above equation is the Mange Elkan distance to measure A and B, where A, the first full name, contains $\{a_1 \ldots a_n\}$ tokens and B, the second full name, contains $\{b_1 \ldots b_m\}$ tokens.

3. ARABIC-TO-ENGLISH MAPPING

There are more than four hundred million Arabic language speakers who live in 22 countries from the Arabian Peninsula across North Africa to the Atlantic Ocean [15]. Arabic is the main language of newspapers, TV news, books, and official works, even though there are different accents across countries and even within countries. In addition, there are more than 1.6 billion Muslims, and most of them have an Arab name. Table 1 shows various spellings of the name "Mohammed" in different nations; there are more than 300 different spellings for the name "Mohammed" alone [2].

This section explains in detail the mapping of each Arabic letter to its closest English letter, as well as their relationships with Soundex groups. There are four groups of Arabic letters that are mapped to English letters:

- Exact match: some of the Arabic letters can be mapped to English letters that have almost the same sound. The EXACT column in Table 3 shows these letters.
- Exact match with two characters: "TH" can be mapped to two Arabic letters (ذ, ث), while the letter "ش" is mapped to either "SH" or "CH."
- Close match: these letters do not have an equivalent sound in English; they are mapped to a character that produces a similar sound (see the Close column in Table 3).
- Diacritic sound: Hamza (ء) and three more diacritics change the sound of some Arabic letters, such as "ا," which can be mapped to A, O, U, E, and I (see the Others column in Table 3).

Table 3. English Arabic Characters Mapping

	EXACT	CLOSE	OTHER
A	ا	ع	ا أ آ ى ء
B	ب		
C			س ش ك
D	د		ض ظ
E			إ ي ئ
F	ف		
G			ج غ
H	ح هـ		
I			إ ى ئ ي
J	ج		
K	ك		خ
L	ل		
M	م		
N	ن		
O			ع و ؤ ا
P			ب
Q		ق	ك
R	ر		
S	س	ص	ش
T	ت	ط	ظ ض ذ ث
U			ع و ا
V			ف
W	و		
X			
Y	ي		
Z	ز		ظ

The authors in [16] demonstrate that Arabic and English lack the corresponding sounds to allow one-to-one matching that ends ambiguity in letter mapping, as the results show in the CLOSE column of Table 3.

The rows in Table 2 are grouped based on Soundex representation, where each row represents a number in the Soundex table. Table 2 has an extra column (Arabic Letters) that represents Arabic letters after they are clustered. Consequently, the results show there is no overlap based on Soundex groups except for the letter "ظ," which is due to some countries' diverse dialects or non-Arabic languages, such as Persian or Urdu.

All letters agree in both languages and have similar enough sounds to generate the same Soundex groups. Therefore, Soundex is the best choice over all the other phonetic algorithms that are English dependent. Adding more groups, such as Refined Soundex, introduces overlap for same Arabic characters; e.g., "ق" (Qaf) points to both K and Q, which are in two different groups in Refined Soundex. This causes a comparison problem for Arabic names such as "Qaddafi" and "Kaddafi," or "Qasim" and "Kasim."

4. PERSONAL IDENTITY MATCHING (PIM) AND PREPROCESSING

The PIM algorithm converts names to points in space that have seven attributes (seven-dimension) representing the groups in the Soundex table, including the ignored characters. The following steps show how we transform Soundex groups to a seven-dimensional vector:

1. Use each row (group) in the Soundex table as an independent group.

2. Add all letters that are ignored by Soundex into the table as group 0 (silent char.).

3. PIM now has 7 groups (0 to 6); refer to Table 2.

4. Assign a value for each character. The difference between two characters ($|\alpha - \beta|$) represents the distance between these characters in the same group. We created a handcrafted model that contains each character value.

5. Create a Vector-Consonant (V_c) with 7 dimensions to hold consonant values, group 1 to group 6, including group 0 if and only if it is the first character in the name. For example: $V_c('\text{Nasser}') = \{null, null, \{S, S\}, null, null, \{N\}, \{R\}\}$; and $V_c('\text{Ali}') = \{A, null, null, null, \{L\}, null, Null\}$.

6. For each attribute in V_c, create a vector of a 3-dimensional representation to hold group 0 characters when they do not appear as the first character. The characters are divided into three groups: "e", "i", and "y" in the first group, "a" and "h" in the second group, and "o", "u", and "y" in the third group. For example, the name "Nasser" is converted to $\{\{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\text{E}, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \text{A}, \phi\}, \{\phi, \phi, \phi\}\}$; and $V_g('\text{Ali}') = \{\{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \text{I}, \phi\}, \{\phi, \phi, \phi\}, \{\phi, \phi, \phi\}\}$ where ϕ is a null value.

Soundex retains the first letter of any name at the beginning of the code, including characters in group-0. In addition, the Jaro-Winkler algorithm improves the Jaro algorithm by boosting the

error score in the prefix characters. The Jaro-Winkler algorithm adds more values with a common prefix at the beginning of the names, up to four letters [12]. The first character(s) of the name is pronounced clearly, and it is rarely spelled incorrectly. Therefore, the PIM algorithm treats each letter in group-0 the same as the letters in the other groups (consonants) if and only if they appear as the first letter. For this reason, there are two values for each character: one value is for the first character of the name, and the second value is for any character except the first character. However, some letters have very close pronunciation even if they appear as the first character of the name, e.g., Auto and Otto. The beauty of PIM is that it can easily manage the similarity between characters. If A and O contain the same value, then they are identical; if the different value between these two characters increases, the similarity decreases (see step 4).

5. PROPOSED ALGORITHM

Each name passes to a function that is called a vector function generator. This function accepts three parameters: Name (in), V_c (out), and V_g (out). After the first name is passed to the function, two vectors (V_{c1}, V_{g1}) represent the first name. The second name is then passed to the same function to produce two other vectors (V_{c2}, V_{g2}). Once the vectors are generated for all names, Euclidean distance (Ed) is used to calculate the distance between the names. The calculation performed as following:

1. Calculate → Ed_C (V_{c1},V_{c2})

2. Calculate → Ed_G (V_{g1},V_{g2})

3. G (Total silent char.) → G = G / 2

4. N (Total char.) N = G + C

5. C (Total consonant) → C = N – G, add more weight to C if silent chars. are not equal in both names

6. Convert both distances to similarities: Sim_C and Sim_G

7. Finally use (2) to calculate PIM similarity

$$Sim_{pim} = \frac{C(Sim_C) + G(Sim_G)}{N} \quad (2)$$

C=Consonant, G=Group-0, & N=Total chars.

To convert a single name to a vector that has seven attributes, the algorithm needs to insert the value of each character into a suitable attribute. This conversion costs $O(|n|)$, where n is the number of characters in the name. If the algorithm compares two names, the complexity is $O(|n|+|m|)$, where n is the length of the first name, and m is the length of the second name. Similarity calculation is constant, as described above. Therefore, total complexity of PIM is $O(|n|+|m|)$. In addition, if there is a list of names, the comparison performance can be improved by storing the vectors of the names in a repository. For future comparison, PIM needs only to convert a given name to a vector and it calculates the similarity between the given name vector

and the vectors in the repository. PIM is used for a single name and cannot understand full name structure. Therefore, we need another algorithm that can deal with such structure.

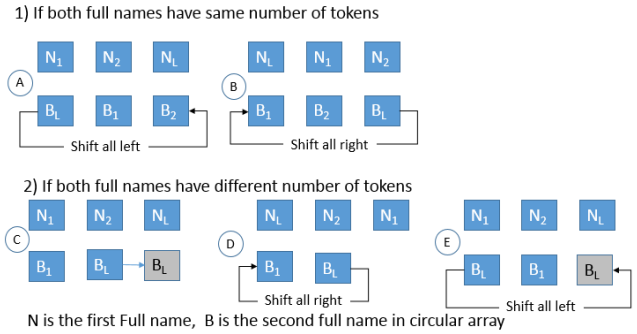


Figure 1. Unmatched Token Adjcement (Names).

We create another algorithm (PIMF) that is an extension of PIM. PIMF comprehends full name structure, as we explained earlier. PIMF accepts the two full names and breaks them into tokens (names). Then, it sends the tokens to PIM to calculate the similarity. PIMF uses the following process:

1. Break both full names into two lists of tokens: $a_1, a_2 \dots , \ a_n$ and $a_1, a_2, \dots , \ a_m$
2. Set FNS = Full name with least tokens
3. For i=1 to FNS Length TotalSum = TotalSum + PIMSimilarity($a_i, b_{i,}, thershold$)
4. If the average of TotalSum > threshold, return TotalSum / FNS length;
5. Return the maximum (Similarity * threshold) based on the comparisons after the following movements:
    a. Shift all tokens in FNS to the left (see Figure. 1 A & E)
    b. Shift all tokens in FNS to the right (see Figure. 1 B & D.)
    c. If the number of tokens is not equal in both names, then
    Move the last token in FNS to the last position of the long name (Figure. 1 C).

PIMF returns a similarity between 1 and threshold for a positive exact match. It also returns a similarity between the threshold and (threshold * threshold) for positive match with token adjustment; otherwise, it returns zero as a negative match.

5. EXPERIMENT

We were provided an excellent dataset by an organization that collects names from many countries. The organization gave us single names only and not full names for privacy and security reasons. The single names are sufficient for our experiment, and we can generate full names from

those names. The dataset contains names and their spelling variations that were entered from different countries, such as the name “Mohammed” in Table 1.

Each instance in the dataset is the tuple of an Arabic name and its equivalent name in English. The total number of tuples in the dataset is that 29,821. Most of the names are Arabic names selected from 21 Arabic countries. These names are written in Arabic modern language, but when the names are transliterated into English, some of them are written based on the pronunciation of dialects. For example, the Arabic name “ذهب” is Romanized as DAHAB, DHAHAB, THAHAB, and THAHB. We checked the dataset manually and inserted a number (Cluster-Id) to each tuple, grouping all English names that referred either to the same Arabic name or to Arabic names that share several names in English. We used the “Cluster-Id” value as a ground truth for our test. In addition, the dataset contains a small fraction of noise (spelling errors) due to the actual data entry. We kept some of this noise to represent natural data entry errors.

We select 2000 names randomly from the dataset. Each of the random names is compared with all of the names in the dataset. Recall and precision are computed for all 2000 names. Then, we measure the F-Measure as our test’s accuracy for 17 algorithms, including PIM. We repeat the test five times with 2000 different random names each time. Each algorithm is tested with a threshold between 0 and 1, and then the best average is recorded.

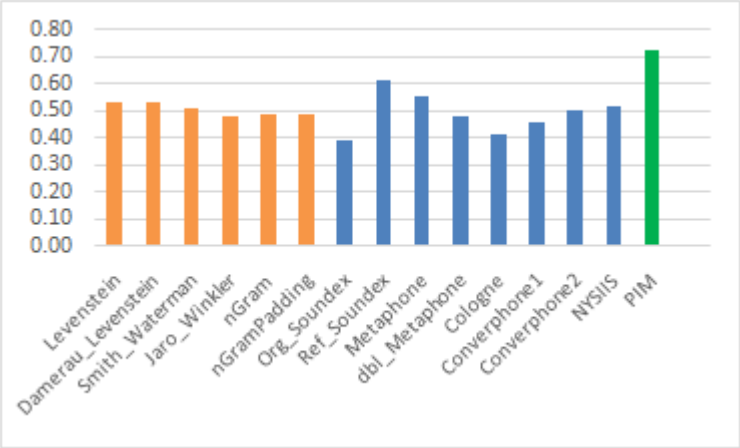


Figure 2**.** F-Measure for all algorithms.

Figure 2 shows the F-Measure for all of the algorithms, including PIM. The results are clustered by colors in Figure 2: Orange represents the pattern match algorithm, and blue represents phonetic algorithms. PIM records the best result with the F-Measure of 0.72, followed by Refined-Soundex, Metaphone, and Levenstein (Edit distance) with F-Measures of 0.61, 0.55, and 0.53, respectively. Soundex has the lowest score 0.39. PIM achieves 11% better accuracy than Refined-Soundex and 33% better accuracy than Soundex. The accuracy of 72% for the dataset that contains names with very similar pronunciation including some noise is considered good performance compared with the17 other algorithms.

Our aim is to identify a person using his/her name. Of course, a person cannot be identified by using a single name, such as a given name or family name only. The full name of a person is needed to be more precise in identification. Whenever a full name is combined with more names, it increases the ability to accurately distinguish that individual. All of the algorithms used in the

first test do not take full name structure into consideration, such as “first-name last-name” or “last-name first-name.” Therefore, we use our PIMF algorithm and the Monge-Elkan algorithm. Both these algorithms use other algorithms that calculate the distance between two names, to compute full name similarity. We use PIM to calculate the similarity between two single names for PIMF and Monger-Elkan because PIM has the best result of all comparable algorithms.

Table 4. Results of the First Experiment

Algorithms	Recall	Precision	F-Measure
Caverphone	0.79	0.33	0.46
Caverphone 2	0.85	0.35	0.50
DamerauLevenstein	0.47	0.62	0.53
Double-Metaphone	0.89	0.34	0.48
Jaro-Winkler	0.40	0.60	0.48
Levenstein	0.46	0.61	0.53
Metaphone	0.73	0.44	0.55
Ngram	0.45	0.54	0.49
NGram-Padding	0.45	0.54	0.49
NYSIIS	0.60	0.46	0.52
PIM	0.73	0.72	**0.72**
Ref-Soundex	0.77	0.51	0.61
SmithWaterman	0. 49	0. 53	0.51
Soundex	0.91	0.25	0.39

We generate a dataset containing full names, each of which was composed of two names (tokens) and four names (tokens). Our assumption is that each full name has the right structure, as was explained earlier. So the first and second tokens can be either a given name or a last name. This dataset contains about 74,866 full names, 25% of which are composed of already existing tokens in different order. The revised order might contravene the full name structure; for example, the name “Ali Nasser Mohammed” cannot match “Nasser Ali Mohammed” but matches “Mohammed Ali”, “Ali Mohamed”, “Ali Nasser”, and “Mohammed Ali Nasser.” This dataset is used to test the performance of the PIMF and Monger-Elkan algorithms.

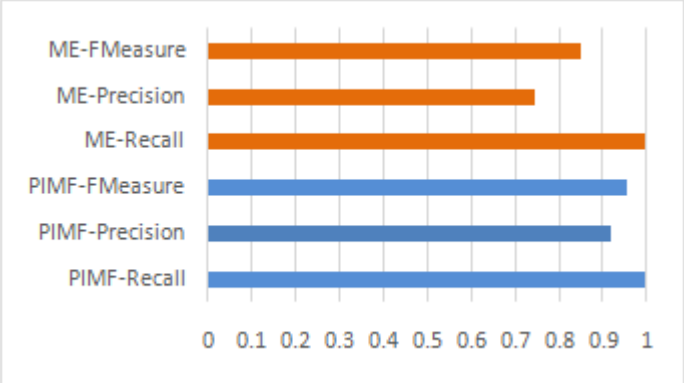


Figure 3. Recall, Precision, and F-Measure for PIMF and Monge-Elkan (ME)

We select 2000 instances randomly from the new dataset, and then the algorithms compare each instance with the entire dataset of 74,866 full names. The F-measure is calculated to check the algorithm performance. This process is repeated five times, and the final results are averaged. Figure 3 displays the recall, precision, and F-Measure for both the PIMF and Monger-Elkan (ME) algorithms. Both algorithms record very high recall, with 0.997 and 0.995 for PIMF and Monger-Elkan, respectively. PIMF has better accuracy than Monger-Elkan with an F-measure of 0.96 versus 0.85. As a result, PIM shows good accuracy at identifying individuals.

Table 5. Results of the Third Experiment

Lang. Match	Algorithm	F-Measure	Recall	Precision
English-Arabic	ASC	0.18	0.23	0.14
	PIM	0.50	0.67	0.40
Arabic-English	ASC	0.23	0.23	0.23
	PIM	0.53	0.61	0.48
Arabic-Arabic	ASoundex	0.78	0.86	0.71
	PIM	0.89	0.84	0.95

The third experiment is to evaluate name comparisons across languages. We implemented PIM to match the names that are written in Arabic with the English alphabet and vice versa. We use the algorithm created by [17,] and label it ASC. This algorithm is built to match personal names in English with names in Arabic script. Table 5 presents the results for PIM and ASC, which score 0.5 and 0.18, respectively, when comparing each English name to the list of Arabic names. When we reverse the test to compare each Arabic name to English names, we get only 3% better accuracy for both PIM and ASC, as shown Table 5. The last test compares names in Arabic to Arabic script. We compare ASoundex, which can be found in [15], with PIM. Table 5 shows the result of ASoundex and PIM for an Arabic-to-Arabic test and end up with F-Measures of 0.78 and 0.89, respectively. Most of the tests show good performance, especially for full names, which are our main targets.

5. CONCLUSION

There are many algorithms for name matching that try to solve the name variation challenges, included name transliteration. Also, cross-language information retrieval cannot deal with OOV words, 50% of which are personal names. Yousef [1] tried to resolve OOV proper names by adding names automatically into a dictionary, and the authors in [8] used probability to transliterate names “on the fly” from English to Arabic.

The experimental results indicate that our proposed algorithms (PIM and PIMF) perform better than most known name-matching algorithms to compare transliterated names and full names. Also, the PIM shows a good outcome when the Arabic language is used as an example to compare names across languages. This comparison was between names that were written in Arabic and English and vice versa. PIMF provides excellent accuracy of up to 96% for full name comparison.

This algorithm can contribute to improving performance in many fields. Security screening systems are one area that can benefit from this algorithm. A given name and a list of names (suspects’ names) can be given as input to this algorithm, which will then return a new list (sub-list) that contains all matched names with a degree of similarity to the given name. This helps to

identify the closest personal names to a given name. In addition, our algorithms contain an excellent environment to accommodate other languages. Cross-language information retrieval can benefit from this algorithm because comparisons can be done "on the fly" for any language that can map its letters to Soundex categories (Table 2) that are converted into a vectors that are standard for all languages.

REFERENCES

[1] Maureen Cooney, "Report Assessing the Impact of the Automatic Selectee and No Fly Lists," 2006.

[2] Christopher Westphal, Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies, 1st ed. CRC Press, Inc. Boca Raton, FL, USA ©2008, 2008.

[3] P. C. P. Christen, "A Comparison of Personal Name Matching: Techniques and Practical Issues," Sixth IEEE Int. Conf. Data Min. - Work., 2006.

[4] Tina Lieu: The Name Matching You Need: A Comparison of Name Matching Technologies, Cambridge, MA (2012).

[5] David Pinto, Darnes Vilariño Ayala, Yuridiana Alemán, Helena Gómez, Nahun Loya, "The Soundex Phonetic Algorithm Revisited for SMS Text Representation," in TSD 2012, 2012, vol. 7499, pp. 47–55.

[6] L. Karl Branting, "Efficient Name Variation Detection", AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection, Arlington, VA, October 13-15, 2006

[7] Ghita Amor-Tijani, "Enhanced english-arabic cross-language information retrieval," George Washington University Washington, DC, USA, 2008.

[8] Larkey, L., AbdulJaleel, N., Connell, M.: What's in a Name?: Proper Names in Arabic Cross Language Information Retrieval. Massachusetts (2003).

[9] K. Darwish, "Named Entity Recognition using Cross-lingual Resources : Arabic as an Example," in Acl, 2013, pp. 1558–1567.

[10] N. AbdulJaleel and L. S. Larkey, "Statistical transliteration for english-arabic cross language information retrieval," in Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03, 2003, p. 139.

[11] Lawrence Philips, "Hanging on the Metaphone", Computer Language, Vol. 7, No. 12 (December), 1990.

[12] W. E. Yancey, "Evaluating string comparator performance for record linkage," Tech. Rep. RR2005/05, US Bureau of the Census 2005.

[13] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive name matching in information integration," IEEE Intell. Syst., vol. 18, no. 5, pp. 16–23, 2003.

[14] M. Sabbagh-nobarian, "The Review of Fields Similarity Estimation Methods," IJMLC Int. J. Mach. Learn. Comput., vol. 2, no. Icmlc, pp. 596–599, 2011.

[15] S. U. Aqeel, S. Beitzel, E. Jensen, D. Grossman, and O. Frieder, "On the development of name search techniques for arabic," J. Am. Soc. Inf. Sci. Technol., vol. 57, no. 6, pp. 728–739, 2006.

[16] A. Lait and B. Randell, "An assessment of name matching algorithms," Newcastle, UK, 1996.

[17] Freeman, S. Condon, and C. Ackerman, "Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance algorithm," … Comput. Linguist., June, pp. 471–478, 2006

AUTHORS

Mazin H. Al-Shuaili is currently a Ph.D. candidate at Florida Institute of Technology (FIT), in Melbourne, FL/USA. In 1998, he graduated in computer science from FIT. In 2000, he obtained his master's degree in software engineering from FIT. His master had focused on software test automation. From 2000 till 2012, he was a system analyst and project manager at Omani government. In May 2016, he is going to graduate and he going back to work with Omani government, Muscat. His research interests are in general areas of Natural language processing (NLP), social network, and data mining.

Marco M. Carvalho is an Associated Professor at the Florida Institute of Technology, in Melbourne, FL/USA. He graduated in Mechanical Engineering at the University Brasilia (UnB – Brazil), where he also completed his M.Sc. in Mechanical Engineering with specialization in dynamic systems. Marco Carvalho also holds a M.Sc. in Computer Science from the University of West Florida and a Ph.D. in Computer Science from Tulane University, with specialization in Machine Learning and Data Mining. At Florida Tech, Dr. Carvalho is the Executive Director of the Harris Institute for Assured Information, and the Principal Investigator of several research projects in the areas of cyber security, information management, networks, and tactical communication systems.

INTENTIONAL BLANK

FEATURE-MODEL-BASED COMMONALITY AND VARIABILITY ANALYSIS FOR VIRTUAL CLUSTER DISK PROVISIONING

Nayun Cho, Mino Ku, Rui Xuhua, and Dugki Min*

Department of Computer, Information & Communications Engineering,
Konkuk University, Seoul, Korea
{nycho,happykus,abealasd,dkmin}@konkuk.ac.kr

ABSTRACT

The rapid growth of networking and storage capacity allows collecting and analyzing massive amount of data by relying increasingly on scalable, flexible, and on-demand provisioned large-scale computing resources. Virtualization is one of the feasible solution to provide large amounts of computational power with dynamic provisioning of underlying computing resources. Typically, distributed scientific applications for analyzing data run on cluster nodes to perform the same task in parallel. However, on-demand virtual disk provisioning for a set of virtual machines, called virtual cluster, is not a trivial task. This paper presents a feature model-based commonality and variability analysis system for virtual cluster disk provisioning to categorize types of virtual disks that should be provisioned. Also, we present an applicable case study to analyze common and variant software features between two different subgroups of the big data processing virtual cluster. Consequently, by using the analysis system, it is possible to provide an ability to accelerate the virtual disk creation process by reducing duplicate software installation activities on a set of virtual disks that need to be provisioned in the same virtual cluster.

KEYWORDS

Virtual Cluster Disk Provisioning, Feature Model-based Virtual Cluster Commonality and Variability Analysis

1. INTRODUCTION

Virtualization is one of the promising solutions to overcome the limitation of computing power using a flexible resource scaling mechanism [1]. There are various researches in this direction to analyze big data with large-scale computational clusters using the virtualization technique [2,3]. Unlike physical clusters, a virtual cluster (VC) has a set of several virtual machines that needs to be provisioned before running on virtualized physical hosts. There are two steps of VC provisioning: VC placement and VC disk provisioning. The VC placement is a key factor to optimize the utilization of virtualized physical hosts using effective scheduling algorithms of underlying computing resources, such as VCPU, memory, network bandwidth, and so on [4,5]. On the other hand, the VC disk provisioning creates a set of virtual disks depending on the demand of the requested virtual cluster. Creating a set of virtual disks is time consuming tasks.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 45–52, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60205

Therefore, the way of VC disk provisioning directly affects the quality of service of cloud provider [6].

The provisioning process starts with the installing system or application software, such as operating systems or middleware, on an empty disk image. Among these software, some of system or application software, are repeatedly requested by users to install the software on virtual disk images. If there are pre-installed virtual disk images in shared storage (e.g., distributed file system), then the virtual disk image can be reused with a cloning mechanism. The cloning method for the virtual disk provisioning significantly reduces time to create a set of virtual disks of a virtual cluster. However, finding cloneable virtual disks that fully meet the demands of software on a virtual cluster is not a trivial task.

In order to find such reusable virtual disks for the virtual cluster disk provisioning, we apply Software product line (SPL) [7] methodology. SPL is a solution to create a collection of similar virtual disk images from existing shared assets (e.g., virtual disk images) by commonality and variability analysis of the product. To describe commonality and variability, this paper employ Feature model (FM) [8] as a metadata of a virtual disk image. FM is a hierarchical representation model that organizes commonality and variability of all the products of the SPL using features and their relationships. Applying FM to a virtual disk image enables disk provisioning system to determine which software features are commonly used in a virtual cluster. However, generating all the FMs related to a virtual cluster in a manual way is a tedious and error-prone effort. Consequently, a commonality-and-variability analyzer is necessary to generate the related FMs of the virtual cluster automatically.

This paper presents a methodology to provisioning a group of virtual disks for a virtual cluster in terms of software product line engineering. The virtual cluster disk provisioning process based on SPL, which includes (1) analyzing common and variant software features of a VC, (2) retrieving reusable virtual disk images, (3) generating virtual disk provisioning plan, and (4) creating virtual disk images. However, among these provisioning phases, this paper focus on the VC commonality and variability analysis with a case study. In order to analyze common and variant software features of a virtual cluster, several functions are needed. Firstly, the basic structure of feature model for a virtual disk image should be defined. Secondly, feature models of virtual disks for a VC should be generated according to the user's requirements. Thirdly, categorizing the type of virtual disk images should be performed automatically for correctness.

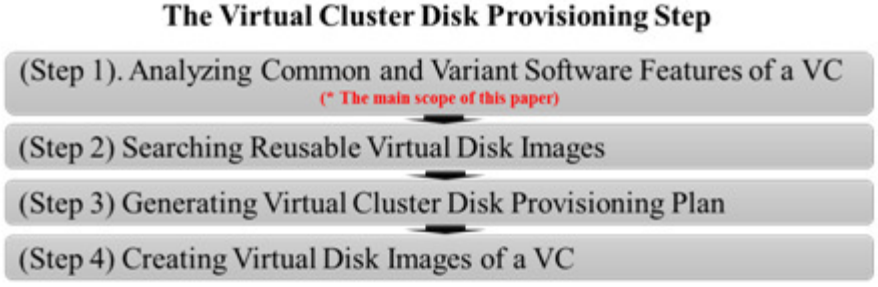


Figure 1. The Disk Provisioning Step of a Virtual Cluster

In this paper, we present the analysis system, named VC C&V analyzer, which archives the aforementioned functional requirements by a feature model reasoning mechanism. At first, VC C&V analyzer generates feature models a VC subgroup by using the VC provisioning specification extended from Open Virtualization Format (OVF). After that, VC C&V analyzer merges the generated feature models to extract the common and variant software features of a given virtual cluster. By using the common and variant software features, VC C&V analyzer generates the final VC commonality and variability feature models of a virtual cluster to classify the types of the disk images that need to be provisioned. Finally, the automated support of the VC C&V analyzer allows to reduce the effort needed to create a set of similar virtual disk images and their similarity investigation.

The remainder of this paper is organized as follows: Section 2 describes the architecture and processing flow of feature model-based VC commonality and variability analyzer with a feasible case study in Section 3. Section 4 discusses related researches and finally, Section 5 presents concluding remarks.

2. VC COMMONALITY AND VARIABILITY ANALYSIS METHOD BASED ON FEATURE MODEL

This section presents a commonality and variability analysis method based on the feature model for virtual cluster disk provisioning. Figure 2 shows a virtual cluster for our case study. Normally, a virtual cluster consists of a set of VC subgroups, such as Hadoop VC subgroup and HBase VC subgroup. Also, each VC subgroup is composed of virtual machines with the same system or application software. In this example, six virtual disk images should be provisioned for a big data processing VC.

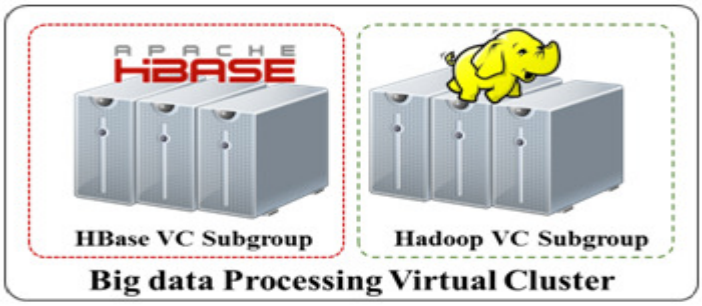


Figure 2. The Big Data Processing Virtual Cluster

In some cases, a set of VC subgroups of a virtual cluster may use similar software platform, such as operating system. For example, Hadoop VC subgroup and HBase VC subgroup may require same system software, named Debian Linux. Thus, commonality and variability analysis among the subgroups of a VC should be done to avoid duplicated installation tasks for the same software platform on a virtual disk. In order to analyze commonality and variability between VC subgroups, the basic structure of feature model for a virtual disk image is needed. Feature model allows the virtual disk provisioning system to categorize types of virtual disks which should be provisioned.

For provisioning the big data processing VC, the virtual disk provisioning system should classify types of virtual disks of Hadoop VC subgroup and HBaseVC subgroup that contain the same

software, such as Debian Linux, or different software, such as hadoop or hbase. Also, the dependencies between software and system architecture, such as AMD64 or i386, should be considered. To support these requirements, we define the basic structure of the feature model that includes architecture, system software, and application software. System software consists of various distributions and each distribution contains its own version. For example, there are several distributions of Linux operating system with version, such as Debian 8.0, Ubuntu 14.04, and CentOS 5.0. According to these types of the distribution, there is a set of variations that determines which packages to be installed in the system distribution, such as minbase, base, buildd, and so on. Similarly, application software consists of name, version, and variants

Since feature model presents commonality and variability of relevant products itself, the VC C&V analyzer generates feature models of VC subgroups using OVF-based virtual cluster specification which defined by the user. The VC C&V analyzer uses the specification as a requirement to meet the needs of a particular purpose of the virtual cluster. This specification involves a virtual hardware specification, such as the number of VCPUs, the size of memory and disks, the virtual network bandwidth required for each virtual machine, and name and version of software of each virtual cluster named *VirtualSystemCollection*. The VC C&V analyzer travels the *VirtualSystemCollections* to extract system and software information from *OperatingSystemSection* and *ProductSection* of VC subgroups. *OperatingSystemSection* involves architecture, distribution, variant, and version of the system software with attributes named id and version. *ProductSection* presents the name and version of application software which the provisioning engine needs to install in a set of virtual disks. Moreover, the end user can describe the relevant software configuration in this section, such as IP address, configurations regarding with a particular application software, and so on.

Using this information of VC subgroups, the VC Subgroup FM Generator (VC Subgroup FMGen) of the VC C&V analyzer maps an architecture variable into the Architecture feature, and name, variant, and version variables into the Distribution and Variant features respectively. Similarly, the VC Subgroup FMGen maps variables of the name, version, and variant of application software to Application feature.

After generating feature models of each VC subgroup, the VC Commonality and Variability(C&V) Feature Model Generator (VC C&V FMGen) merges the generated feature models to analyze the commonality and variability between VC subgroups. If there is only one VC subgroup in the provisioning specification, the VC C&V analyzer skips this step. In our research, FAMILIAR [9] is used for this merging step. In other words, the core features of the merged feature model can be interpreted as common features between VC subgroups, whereas different features can be defined as the various features of the VC subgroups. Using “merge” function of the FAMILIAR, VC C&V FMGen recursively combines the feature models of VC subgroups and generates a VC C&V feature model of the whole VC.

The generator performs “cores” and “mergeDiff” functions to divide the merged feature model into the VC commonality feature model and the VC variability feature model. Some of these feature models contain more than one product (e.g., virtual disks) and some of them involve an individual product of a virtual disk. Since we employ a feature model to describe the metadata of each virtual disk, the generator splits the model to produce a specific type of virtual disk in a case of the feature models that include more than one product. Consequently, VC C&V analyzer

generates final results including a list of locations of the generated feature models and the number of related VC subgroups of each feature model.

3. CASE STUDY

This section presents a case study of applying the proposed VC C&V analysis method to a big data processing VC. Basically, Hadoop and HBase requires master and slave nodes to handle big data in a distributed way. To avoid the performance degradation under hot spotted case [10], we separately design VC subgroups along with the big data processing middleware (e.g., Hadoop and HBase). Figure 3 shows the detailed description of a virtual cluster including Hadoop and HBase subgroups.From now, this section describes four processing steps to analyze commonality and variability of a big data processing VC by using the VC C&V analyzer.

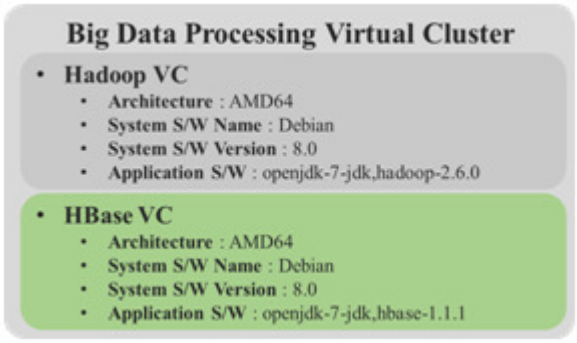


Figure 3. A Case Study of the Big Data Processing Virtual Cluster

Step 1 – Extracting architecture, name, variant and version of the system and application software: Firstly, the VC C&V Analyzer extracts the architecture, name, version, and variant of system and application software of the big data processing virtual cluster as shown in Table 1. The result can be categorized by the name of each VC subgroup. Consequently, the extracted information of system and application software is imported into the VC subgroup Feature Model Generator (VC Subgroup FMGen).

Table 1. The extracted information of system and application software of each VC subgroup.

VC Subgroup Name	Software Type	Software Name	Software Version	Software Variant	Architecture
Hadoop VC	System Software	Debian	8.0	base	amd64
	Application Software	Openjdk	7.0	-	amd64
	Application Software	Hadoop	2.6.0	-	amd64
HBase VC	System Software	Debian	8.0	base	amd64
	Application Software	Openjdk	7.0	-	amd64
	Application Software	HBase	1.1.1	-	amd64

Step 2 – Generating VC subgroup feature models: The VC subgroup Feature Model Generator in the VC C&V analyzer maps the resulting information of system and application software to the basic structure of feature model. Figure 4 shows a snippet of the generated Hadoop and HBase VC subgroup feature models. By using these feature models, VC C&V Analyzer performs the

feature model comparison to determine which software features are commonly used among the VC subgroups.

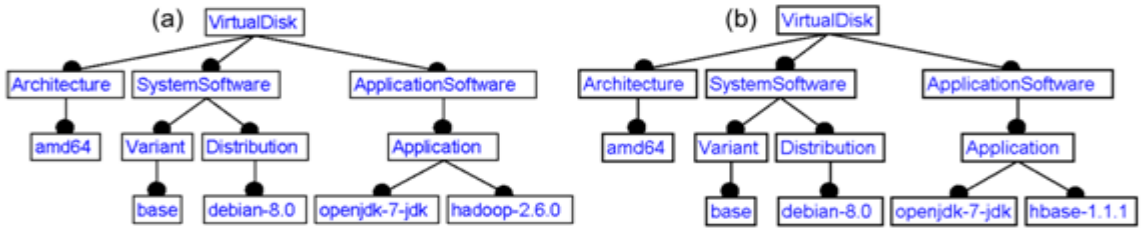


Figure 4. The generated feature models of Hadoop (a) and HBase(b) VC subgroups

Step 3 – Comparing feature models of each VC Subgroup: It is easy to categorize commonality and variability of the resulting feature models in manual. However, to automate such analysis activities, we employ a method of feature model reasoning using FAMILIAR framework. Figure 5 shows the consequent results of step 3.

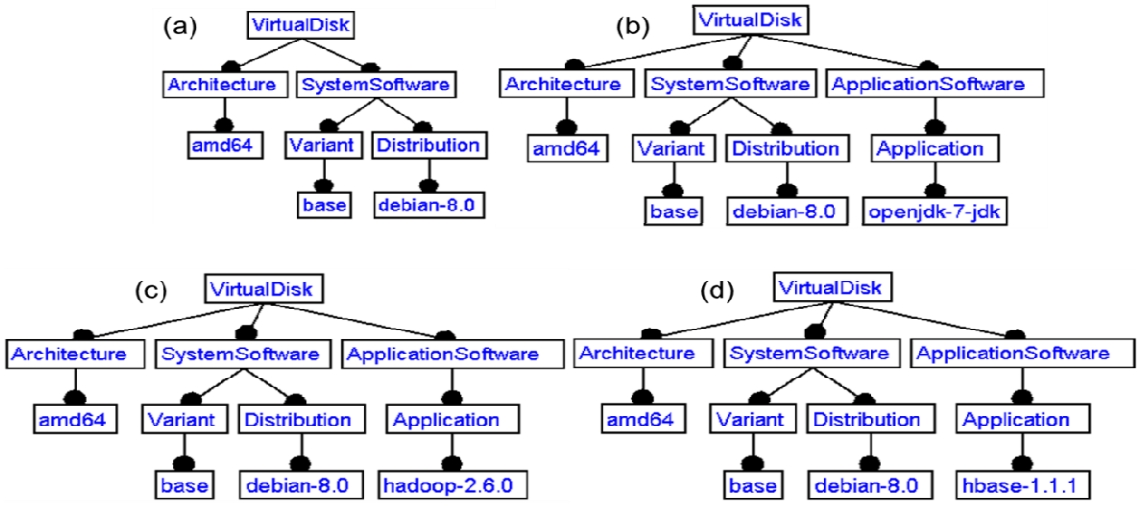


Figure 5. The comparison results between Hadoop and HBase VC subgroups.
5(a) and 5(b) presents common software features, and 5(c) and 5(d) indicates different software features among the VC subgroups.

As shown in Figure 5, VC C&V Analyzer generates feature models of commonality and variability based on Hadoop and Hbase VC subgroups. Also, each feature model indicates an individual type of virtual disk, such as amd64 architecture based debian-8.0, openjdk-7-jdk, hadoop-2.6.0, and hbase-1.1.1. These results will be used on the next virtual disk provisioning step to investigate a reusable virtual disk from the reusable asset repository.

Step 4 – Generating final results of the VC C&V Analysis: As a final result of the VC C&V analyzer, we employ a JSON-based result model. The final result is generated by two steps. Firstly, it categorizes a set of groups which use shared virtual disks among the virtual machines. In this case study, we design none of VC subgroups share virtual disks with other virtual machines. Secondly, it describes required quantity and location of the commonality and variability feature models. From the final result, it is easy to determine which reusable virtual disk meets the analyzed types of virtual disks or not.

4. RELATED WORK

This section presents some efforts in the area of virtual disk image provisioning in context of SPL [7, 11, 12, 13]. Among these researches, Wittern, Erik, et al. [11] present an Infrastructure-as-a-Service (IaaS) deploy model to describe IaaS consumer requirements, including VMs, virtual disk images, and software installed on the images using feature model. Once IaaS consumer selects the cloud provider, VM type, and virtual disk image for the VM, deployment engine invokes a web service call to instantiate VM described in the selected IaaS deploy model. After instantiate VM, the software installation tasks are executed via SSH using a configuration management tool, such as Chef. Also, Dougherty, et al. [12] shows an approach to optimizing configuration and cost of auto-scaling cloud infrastructure. They provide a feature model of virtual machine configuration that captures software platform, including operating system and applications.

Similar to the aforementioned research, the configuration of cloud infrastructure is generated by a selection of features from the feature model in a manual way. Using the configuration, they aim to find a matching virtual machine that already pre-booted in the auto-scaling queue. Krsul, Ivan, et al. [13] provides a direct acyclic graph-based model for configuration activities of a VM. If there a partial graph matching with a set of graphs stored in the repository, named VM shop, the system configures the partial matches of cache VM images as follow as the production line which controls procedures for cloning and configuring a VM. There are several works to employ SPL to create images for a virtual machine, however, none of the works has been addressed how effectively SPL can be used for provisioning virtual disk images of a virtual cluster.

5. CONCLUSION

This paper described a way to provisioning virtual disk images of a virtual cluster via feature model-based VC description and their commonality analysis. We presented our methodology in the context of a feature model-based commonality and variability analysis of a VC that provides an ability to accelerate the provisioning process by reducing duplicate type of virtual disks in the same virtual cluster.

We presented detailed processing flow of the VC C&V Analyzer to determine which types of virtual disks should be provisioned together in a given virtual cluster. We have applied VC C&V Analyzer to investigate common and variant types of virtual disk images among VC subgroups of big data processing. Moreover, our experience in using the VC C&V Analyzer to generate a feature model of each VC subgroup and compare these resulting feature models to determine software which need to be provisioned commonly in this case study. There are still remaining important issues concerning VC disk creation by using the result of VC C&V analysis. We are addressing these remaining challenges as part of our future work.

ACKNOWLEDGEMENTS

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1011) supervised by the IITP(Institute for Information & communications Technology Promotion).

REFERENCES

[1] Foster, I., Zhao, Y., Raicu, I., & Lu, S., (2008) "Cloud computing and grid computing 360-degree compared. In Grid Computing Environments", Workshop, GCE'08, pp. 1-10. IEEE.

[2] Foster, I., Freeman, T., Keahy, K., Scheftner, D., Sotomayer, B., & Zhang, X., (2006) "Virtual clusters for grid communities. In Cluster Computing and the Grid", CCGRID 06. Sixth IEEE International Symposium on Vol. 1, pp. 513-520. IEEE.

[3] Hoffa, C., Mehta, G., Freeman, T., Deelman, E., Keahey, K., Berriman, B., & Good, J., . (2008) "On the use of cloud computing for scientific workflows" In: eScience, eScience'08. IEEE Fourth International Conference on, pp. 640-645. IEEE.

[4] Sotomayor, B., Montero, R. S., Llorente, I. M., & Foster, I, (2009) "Virtual infrastructure management in private and hybrid clouds. Internet computing", Vol. 13, No. 5, pp. 14-22. IEEE.

[5] Mino Ku, (2015) "Flexible and Extensible Framework for Virtual Cluster Scheduling", Ph.D. Thesis, Konkuk University.

[6] Juve, G., & Deelman, E.: Wrangler, (2011) "Virtual cluster provisioning for the cloud", In Proceedings of the 20th international symposium on High performance distributed computing, pp. 277-278. ACM.

[7] Pohl, K., Böckle, G., & van der Linden, F. J, (2005) "Software product line engineering: foundations, principles and techniques", Springer Science & Business Media.

[8] Kang, K. C., Lee, J., & Donohoe, P, (2002) "Feature-oriented product line engineering", IEEE software, Vol. 4, pp. 58-65, IEEE.

[9] Acher, M., Collet, P., Lahire, P., & France, R. B, (2013) "Familiar: A domain-specific language for large scale management of feature models", Science of Computer Programming, Vol. 78, No. 6, pp. 657-681.

[10] Wu, Y., & Gong, G, (20 A Fully Distributed Collection Technology for Mass Simulation Data. In Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on, pp. 1679-1683. IEEE. (2013)

[11] Wittern, E., Lenk, A., Bartenbach, S., & Braeuer, T, (2014) "Feature-based Configuration of Vendor-independent Deployments on IaaS", In Enterprise Distributed Object Computing Conference (EDOC), IEEE 18th International, pp. 128-135. IEEE.

[12] Dougherty, B., White, J., & Schmidt, D. C, (2012) "Model-driven auto-scaling of green cloud computing infrastructure", Future Generation Computer Systems, Vol. 28, No. 2, pp. 371-378.

[13] Krsul, I., Ganguly, A., Zhang, J., Fortes, J. A., & Figueiredo, R. J, (2004) "Vmplants: Providing and managing virtual machine execution environments for grid computing", In Supercomputing, 2004. Proceedings of the ACM/IEEE SC2004 Conference, pp. 7-7. IEEE.

COMPARISON OF OPEN-SOURCE PAAS ARCHITECTURAL COMPONENTS

Mohan Krishna Varma Nandimandalam[1] and Eunmi Choi[2]

[1]Graduate School of Business IT, Kookmin University, Seoul, South Korea
nmohankv@kookmin.ac.kr
[2]Corresponding Author, School of Business IT,
Kookmin University, Seoul, Korea
emchoi@kookmin.ac.kr

ABSTRACT

Cloud computing is a widely used technology with three basic service models such as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). This paper focuses on the PaaS model. Open source PaaS model provides choice of cloud, developer framework and application service. In this paper detailed study of four open PaaS packages such as AppScale, Cloud Foundry, Cloudify, and OpenShift are explained with the considerable architectural component aspects. We also explained some other PaaS packages like Stratos, Stakato and mOSAIC briefly. In this paper we present the comparative study of major open PaaS packages.

KEYWORDS

Cloud Computing, AppScale, Cloud Foundry, Cloudify, OpenShift, Stackato & Stratos

1. INTRODUCTION

Cloud computing is an emerging paradigm in which computers are networked to provide storage and compute services using virtualization technology. Cloud computing must satisfy five essential characteristics. They are on demand service, access network, resource pooling, elasticity and measured services. To achieve these five essential characteristics, cloud computing provides three kinds of service models: Software as a Service (SaaS), Platform as a Service (PaaS) [7] and Infrastructure as a Service (IaaS) [8]. Cloud computing service models are shown in Figure 1. CRM applications are widely used services in the SaaS. Application platform delivered as a service is described as PaaS and it is used to deploy the user code. AppScale [2], Cloud Foundry, Cloudify and OpenShift open-source environments can be used as PaaS. IaaS is used to build their private infrastructure, which reduces the setup cost. IaaS can provide virtualized resources such as computation, storage and communication. Eucalyptus [1], open stack and cloud stack open-sources can be used to provide IaaS.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 53–62, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60206

Figure 1. Cloud computing service models

This paper will focus on the PaaS service model. It is easy to deploy, run and scale application using PaaS. Some of the PaaS have limited language and framework support. They do not deliver key application services needed for cloud applications. They sometime restrict deployment to a single cloud. Whereas open PaaS provides choice of cloud like private, public or hybrid, choice of developer framework like spring, ruby, or java and application services like mongoDB, MySQL, or PostgreSQL for running our applications. This paper deals with the architectural components of major open PaaS packages like AppScale, Cloud Foundry, Cloudify and OpenShift.

The paper is organized as follows. Section 2 introduce AppScale and its components, Cloud Foundry architecture and component explanation given in Section 3, Cloudify open PaaS is explained in Section 4, Section 5 deals with OpenShift, other open PaaS technologies are introduced in Section 6, comparison of open-source PaaS technologies are given in Section 7 and finally Section 8 concludes the paper.

2. APPSCALE

AppScale [3] is a scalable, distributed, and fault-tolerant cloud runtime system that executes over cluster resources. It can be deployed on Xen [4], Kernel-based Virtual Machine (KVM), Amazon EC2 or Eucalyptus. AppScale initial design utilizes the standard three-tier web deployment model in the design. In the later design cycles more components are added to the AppScale. Table 1 shows the AppScale components, language used to design the component and their functionality.

Table 1. AppScale Components

Component	Language	Functionality
AppController	Ruby	Executes on every node and starts automatically when the guest virtual machine boots
AppLoadBalancer	Ruby on Rails	Processes arriving requests from users and forwards them to the application server
AppServer	Python	Running through a number of distant hosts to support automated execution of applications
Database Master	Python	Offers persistent storage for applications, processes protocol buffers from apps and makes requests on its behalf to read and write data to the data store
Database Slave	Python	Facilitate distributed, scalable, and fault tolerant data management
AppScale Tools	Ruby	Starts an AppScale system, deploys and tear down applications, queries the state and performance of AppScale deployment or application, and manipulates AppScale configuration and state

3. CLOUD FOUNDRY

Cloud Foundry [10] is an open PaaS, which provides choice of clouds, developer frameworks and application services. Cloud Foundry makes application development faster and easier. We can build, test, deploy and scale applications with help of Cloud Foundry. It is an open-source project available through a variety of private cloud distributions and public cloud instances. Cloud Foundry started as a platform to deploy Java Spring applications on Amazon Web Services. VMware acquired the Cloud Foundry and made it into an open-source, multi-language and multi-framework PaaS. Cloud Foundry supports multiple languages and multiple runtimes such as Java, Ruby, Scala, spring and Node.js. Cloud Foundry can run on anything like laptop, desktop, micro cloud, private cloud or public cloud. So, it is called as open PaaS as shown in Figure 2. Cloud Foundry has three dimensions to the platform: choice of frameworks, choice of application services and the deployment choice. Cloud Foundry supports spring for Java, Rails and Sinatra for Ruby, Node.js and JVM languages like Groovy, Grails and Scala. It also supports Microsoft .NET Framework and became the first non-Microsoft platform to support .NET.

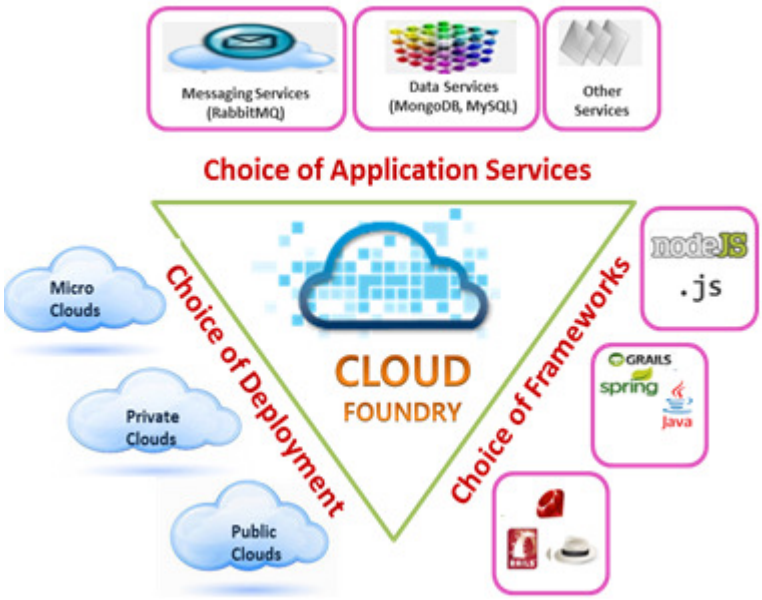


Figure 2. Cloud Foundry as Open PaaS

Cloud Foundry supports RabbitMQ for messaging, MongoDB and Redis for NoSQL, relational databases MySQL and PostgreSQL. Cloud Foundry can be deployed on notebooks through Micro Cloud Foundry. It is the complete version of Cloud Foundry designed to run in a virtual machine. It can also be deployed on Private Cloud or Public Cloud. These features made Cloud Foundry as a flexible PaaS.

Cloud Foundry components perform routing, authentication, messaging, logging, application storage and execution, provide services and take care of application life cycle. The router routes incoming traffic to the appropriate component, usually the Cloud Controller or a running application on a DEA (Droplet Execution Agent) node. The User Account and Authentication (UAA) server work with Login Server to provide identity and authentication management. OAuth2 Server is uses as the user account and authentication server. Cloud controller and health

manager components take care of the application lifecycle in the cloud foundry. Cloud controller is responsible for managing the lifecycle of applications. When a developer pushes an application to cloud foundry, application is targeting the cloud controller. Cloud controller then stores the raw application bits, creates a record to track the application metadata, and directs a DEA node to stage and run the application. Health manager monitor applications to determine their state, version, and number of instances. Applications state may be running, stopped, or crashed. Health manager determine applications expected state, version, and number of instances. It reconciles the actual state of applications with their expected state. Health manager directs the cloud controller to take action to correct any discrepancies in the state of applications. The Droplet Execution Agent manages application instances, tracks, started instances, and broadcasts state messages. Application instances live inside warden containers. Containerization ensures that application instances run in isolation, get their fair share of resources, and are protected from noisy neighbours. Blob Store holds the application code, build packs, and droplets. Applications typically depend on services like databases or third-party SaaS providers. When a developer provisions and binds a service to an application, the service broker for that service is responsible for providing the service instance. Cloud Foundry uses a lightweight publish-subscribe and distributed queueing messaging system for internal communication between components. This internal communication performed via message bus. The metrics collector gathers metrics from the components. Operators can use this information to monitor an instance of Cloud Foundry. The application logging aggregator streams the application logs to the corresponding developers. Cloud Foundry components are shown in Figure 3.

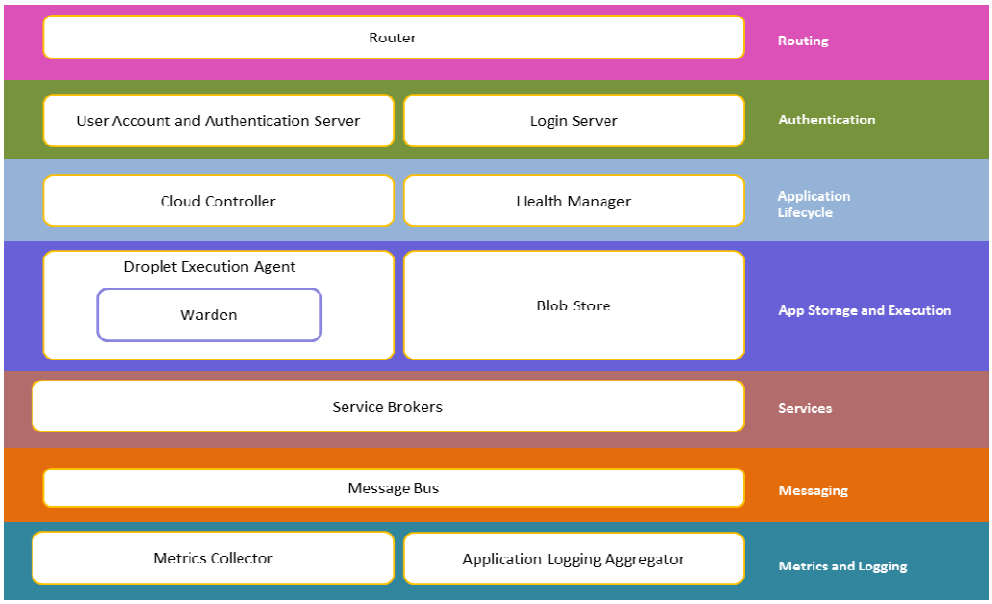


Figure 3. Cloud foundry components

4. CLOUDIFY

Cloudify [11] is another open PaaS cloud application manager. It automates common processes needed to perform and to manage the applications in a cloud environment. Cloudify composed of three main components. The components are Command line interface client, Agents, and Manager. Command line interface client is an executable file which is written in Python. It is

packaged with python and relevant dependencies in an executable file. Command line interface client can run on Windows, Linux and Mac operating systems. Command line interface client preform two tasks. First one is manager bootstrapping and another is managing applications. Bootstrapping is the process of installing the Cloudify manager. Command line interface client provides the user with the full set of functions for deploying and managing applications including log and event browsing.

Cloudify Agents are responsible for managing the manager's command execution using a set of plugins. There is a manager side agent per application deployment and optional agent on each application Virtual Machine (VM). The manager side agents handle IaaS related tasks, like creating a VM or a network, and binding a floating IP to a VM. Manager side agents can also be used with other tools such as REST to remotely execute tasks. The application side agents are optionally located on application VM's. The user can state in the blueprint which VM's will have an agent installed on them. The application side agents are installed by the manager side agent as part of the VM creation task. Once running, the application side agent can install plugins and execute tasks locally. Typical tasks will be middleware installation and configuration, and application modules deployment.

Cloudify Manager deploys and manages applications described in blueprints. The manager's main responsibilities are to run automation processes described in workflow scripts and issue execution commands to the agents. Cloudify is controlled via a REST API. The REST API covers all the cloud orchestration and management functions. Cloudify's Web GUI works with the REST API to add additional value and visibility. Cloudify uses a Workflow engine to allow automation process through built-in and custom workflows. Workflow engine is responsible of timing and orchestrating tasks for creating or manipulating the application components. The user can write custom workflows in Python using API's that provide access to the topology components.

Figure 4. Cloudify Stack

Cloudify uses different databases as data store, some of the technologies for processing and messaging, and different servers as front end. Total stack is shown in Figure 4. Cloudify uses elastic search as its data store for deployment state. The deployment model and runtime data are stored as JSON documents. Blueprints are stored in the elastic search and it is used as runtime DB. Cloudify uses InfluxDB as the monitoring metrics repository. Influx provides flexible schema for metrics and metrics metadata as well as a query language. Cloudify stores every metric reported by a monitoring tool into influxdb and define time based aggregations as well as

statistic calculations. Clodify uses RabbitMQ task broker for messaging. Cloudify offers a policy engine that runs custom policies in order to make runtime decisions about availability, service level agreement, etc. For example, during installation, the policy engine consumes streams of events coming from monitoring probes or tools. The policy engine analyses these streams to decide if a specific node is up and running and provides the required functionality. Policies are registered, activated, deactivated and deleted by the Workflow Engine. For logging purpose logstash is used and agent play main role in processing. Nginx proxy and file server, Flask or Gunicorn REST server, and Node.js GUI servers can be used as front end in the Cloudify.

5. OPEN SHIFT

OpenShift [12] enables us to create, deploy and manage applications within the cloud. Two basic functional units of the Openshift are the Broker and Node servers. Communication between the Broker and Nodes is done through a message queuing service. Broker is the single point of contact for all application management activities. It is responsible for managing user logins, DNS, application state, and general orchestration of the applications. Customers don't contact the broker directly; instead they use the Web console or CLI tools to interact with Broker over a REST based API. Nodes are the systems that host user applications. In order to do this, the Nodes are configured with Gears and Cartridges. A gear represents the part of the Node's CPU, RAM and base storage that is made available to each application. An application can never use more of these resources allocated to the gear, with the exception of storage. OpenShift supports multiple gear configurations, enabling users to choose from the various gear sizes at application setup time. When an application is created, the Broker instructs a Node to create a new gear to contain the application. Cartridges represent pluggable components that can be combined within a single application. These include programming languages, database engines, and various management tools. Users can choose from built-in cartridges that are served directly through OpenShift, or from community cartridges that can be imported from a git repository. The built-in cartridges require the associated languages and database engines to be installed on every Node.

6. OTHER PAAS

In this section we are going to give brief introduction about Stratos, Stakato and mOSAIC open PaaS environments.

6.1. Stratos

Apache Stratos [5] is a highly-extensible PaaS framework that helps to run Apache Tomcat, PHP, and MySQL applications, and can be extended to support many more environments on all major cloud infrastructures. For developers, Stratos provides a cloud-based environment for developing, testing, and running scalable applications. In Single JVM deployment model Stratos could accommodate up to 100 cartridge instances. In a distributed deployment model Stratos could accommodate up to 1000 cartridge instances.

6.2. Stakato

Stackato [6] is open PaaS software based on Cloud Foundry, Docker and other open-source components. It has multi-tenancy capabilities and can be installed on internal infrastructure or public cloud. Multi-tenancy capabilities are important because they allow us to run multiple

applications on the same IaaS infrastructure. Stackato allows developers to automatically package applications into their own Docker containers and scales instances up or down on demand. Stackato provisions all required components, including languages, frameworks and service bindings, automates logging and monitoring, allows for automated application versioning and rollback.

6.3. mOSAIC

mOSAIC [9] is an open-source API and platform for designing and developing multi-Cloud-oriented applications. The architecture has been designed with open and standard inter faces. The main goal is to provide a unified cloud programming interface which enables flexibility to build applications across different cloud providers. The main middleware components providing integration features are the Cloudlet, Connector, Interoperability, and Driver API. The Cloudlet and Connector API layers facilitate the integration into the target language environment which is used by the developers in their applications. The Driver API layer provides abstraction over resource allocation on top of the native resource API. Interoperability API is the middleware layer that integrates the connector API and compatible driver API implementations that could be written in different languages. It is a remote API that follows the model of RPC with functionalities including marshalling, request/response correlation, and error detection. Apart from its cloud integration features, mOSAIC framework is promised to have a semantic-oriented ontology for describing cloud resources.

7. COMPARISON OF MAJOR PAAS

This section compares the major open PaaS frameworks. Table 2 shows the basic functionality and its corresponding AppScale, Cloud Foundry, Cloudify, and OpenShift architectural components.

Table 2. Open PaaS Components comparison

Functionality	**AppScale**	**Cloud Foundry**	**Cloudify**	**OpenShift**
Core functionality	AppController	Cloud controller	Manager	Broker
Providing third party database services	Database Master	Service Broker	Agent	Cartridge
Routing of incoming traffic	AppLoadBalancer	Router	Manager	REST API
Querying the state of apps	AppScale Tools	Cloud controller	CLI client	Broker
Messaging	AppController	Message Bus	Manager	Broker
Application instance management	AppServer	Droplet Execution Agent	Agent	Node
Application state change	AppLoadBalancer	Health Manager	Manager	Broker
Containerization	Database Slave	Warden	Agent	Gear
Load balancing of user requests	AppLoadBalancer	Droplet Execution Agent	Manager	Broker
Framework provider	AppServer	Blob Store	Agent	Cartridge

Table 3 shows the AppScale, Cloud Foundry, Cloudify, and OpenShift PaaS supported languages (java, python, ruby), databases (MongoDB, MySQL, HBase) and frameworks (spring, rails, and flask). In OpenShift, languages and databases are supported in the form of cartridges. User defined cartridges are also allowed in OpenShift. Cloud Foundry provisions languages in the

form of build packs. Users can also pick to write their own build packs. Cloudify, Cloud Foundry and Openshift have extensible language support feature.

Table 3. Language, Database and Frameworks supported by open PaaS

	Languages	**Databases**	**Frameworks**
AppScale	Python, Java, Go, PHP	Cassandra, HBase, Hypertable, MongoDB, SimpleDB, MySQL	Django, Flask, Spring
Cloud Foundry	Java, Ruby, Scala, Node.js, Groovy, Grails, PHP, Go, Python	MonogoDB, MySQL, PostgreSQL	Spring, Rails, Grails, Play, Sinatra
Cloudify	Java, PHP, Ruby	MySQL, MongoDB	-
OpenShift	Java, PHP, Ruby, Python, Perl, JavaScript, Node.js	PostgreSQL, MySQL, MongoDB	Rails, Flask, Django, Drupal, Vert.x

Table 4 shows the features support by AppScale, Cloud Foundry, Cloudify, and OpenShift platforms.

Table 4. Open PaaS Considerable Feature Support

Features	**AppScale**	**Cloud Foundry**	**Cloudify**	**OpenShift**
Relational database support	Yes	Yes	Yes	Yes
NoSQL database support	Yes	Yes	Yes	Yes
Horizontal Scaling	Yes	Yes	Yes	Yes
Vertical Scaling	No	Yes	No	Yes
Auto Scaling	Yes	No	Yes	Yes
Spring Framework support	Yes	Yes	No	No

8. CONCLUSIONS

Cloud computing service models like Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) are introduced in this paper. PaaS is explained in detail with the help of open PaaS packages like AppScale, Cloud Foundry, Cloudify, and OpenShift. AppScale components are explained in table format, Cloud Foundry components are explained in detailed with a diagram, Cloudify and OpenShift components are also explained. Stakato, Stratos and mOSAIC open PaaS environments also explained in this paper. Comparative study is performed among the AppScale, Cloud Foundry, Cloudify and OpenShift open PaaS componets.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education. (Grant Number: 2011-0011507).

REFERENCES

[1] Daniel Nurmi, Richard Wolski, Chris Grzegorczyk, Graziano Obertelli, Sunil Soman, Lamia Youseff & Dmitrii Zagorodnoy, (2009) "The Eucalyptus Open-Source Cloud-Computing System", CCGrid, pp124-131.

[2] Bunch, C., Chohan, N., Krintz, C., Chohan, J., Kupferman, J. & Lakhina, P., et al., (2010) "An Evaluation of Distributed Datastores Using the AppScale Cloud Platform", IEEE International Conference on Cloud Computing.

[3] Bunch, Chris, Navraj Chohan & Chandra Krintz, (2011) "Appscale: open-source platform-as-a-service", UCSB Technical Report.

[4] Varma, N. M. K., Min, D. & Choi, E. (2011) "Diagnosing CPU utilization in the Xen virtual machine environment", In Computer Sciences and Convergence Information Technology (ICCIT), 6th International Conference, pp. 58-63, IEEE.

[5] Pawluk, P., Simmons, B., Smit, M., Litoiu, M. & Mankovski, S. (2012) "Introducing STRATOS: A cloud broker service", In 2012 IEEE Fifth International Conference on Cloud Computing, pp. 891-898, IEEE.

[6] Fortiş, T. F., Munteanu, V. I., & Negru, V., (2012), "Towards a service friendly cloud ecosystem", In Parallel and Distributed Computing (ISPDC), 11th International Symposium, pp. 172-179, IEEE.

[7] Hossny, E., Khattab, S., Omara, F. & Hassan, H., (2013) "A Case Study for Deploying Applications on Heterogeneous PaaS Platforms", In Cloud Computing and Big Data (CloudCom-Asia), International Conference on (pp. 246-253), IEEE.

[8] Varma, N. M. K. & Choi, E., (2013) "Extending Grid Infrastructure Using Cloud Computing", In Ubiquitous Information Technologies and Applications, pp. 507-516, Springer Netherlands.

[9] Marpaung, J., Sain, M. & Lee, H. J., (2013) "Survey on middleware systems in cloud computing integration", In Advanced Communication Technology (ICACT), 15th International Conference, pp. 709-712, IEEE.

[10] D. Bernstein, (2014) "Cloud Foundry Aims to Become the OpenStack of PaaS", IEEE Cloud Computing, (2), 57-60.

[11] Graham, S. T. & Liu, X., (2014) "Critical evaluation on jClouds and Cloudify abstract APIs against EC2, Azure and HP-Cloud", In Computer Software and Applications Conference Workshops (COMPSACW), IEEE 38th International, pp. 510-515, IEEE.

[12] A. Lomov, (2014) "OpenShift and Cloud Foundry PaaS: High-level Overview of Features and Architectures", Available at www.altoros.com/openshift_and_cloud_foundry_paas.html.

AUTHORS

Name: **Mohan Krishna Varma Nandimandalam**
Address: B-304, DIS Lab, School of Business IT, International Building,
Kookmin University, Seoul-136702, South Korea
Education: Completed Bachelor of Computer Applications degree in 2002, Received Master of Science in Information Systems degree in 2004 and Master of Technology in Computer Science and Engineering in 2007 from VIT University, India. At present studying Ph.D. in Graduate School of Business IT, Kookmin University, South Korea.

Eunmi Choi is a Professor in the School of Business IT,
Chairperson of School of Management Information Systems,
Head of Distributed Information System & Cloud Computing Lab., and
Executive Chief of Business IT Graduate School at Kookmin University, Korea,
Her current research interests include big data infra system and processing, cloud computing, cyber physical system, information security, distributed system, SW meta-modelling, and grid & cluster computing. Professor Choi received and MS and PhD in computer science from Michigan State University, U.S.A., in 1991 and 1997, respectively, and BS in computer science from Korea University in 1988.

A CLOUD BROKER APPROACH WITH QOS ATTENDANCE AND SOA FOR HYBRID CLOUD COMPUTING ENVIRONMENTS

Mário Henrique de Souza Pardo, Adriana Molina Centurion,
Paulo Sérgio Franco Eustáquio, Regina Helena Carlucci Santana,
Sarita Mazzini Bruschi and Marcos José Santana

Institute of Mathematical and Computer Sciences (ICMC)
University of São Paulo (USP)
São Carlos, Brazil
{mhpardo,amolina,psfe,rcs,sarita,mjs}@icmc.usp.br

ABSTRACT

Cloud Computing is the industry whose demand has been growing continuously since its appearance as a solution that offers different types of computing resources as a service over the Internet. The number of cloud computing providers grows into a run, while the end user is currently in the position of having many pricing options, distinct features and performance for the same required service. This work is inserted in the cloud computing task scheduling research field to hybrid cloud environments with service-oriented architecture (SOA), dynamic allocation and control of services and QoS requirements attendance. Therefore, it is proposed the QBroker Architecture, representing a cloud broker with trading features that implement the intermediation services, defined by the NIST Cloud Computing Reference Model. An experimental design was created in order to demonstrate compliance to the QoS requirement of maximum task execution time, the differentiation of services and dynamic allocation of services. The experimental results obtained by simulation with CloudSim prove that QBroker has the necessary requirements to provide QoS improvement in hybrid cloud computing environments based on SOA.

KEYWORDS

Cloud Broker, Cloud Computing, SOA, QoS, dynamic service allocation, deadline, task scheduling algorithm, intermediation, NIST Reference Model.

1. INTRODUCTION

The growing adoption of cloud computing as a solution to infrastructure, platform or software offering as a service has grown so much (about 32.8% increase, according to a forecast by the Gartner Group [1] for the year 2015) that the market and the cloud computing environments are becoming increasingly crowded and complex.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 63–82, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60207

This complexity goes beyond the physical infrastructure of data centers, as currently the major trend has been the multiplicity of providers and the construction of complex organizations involving multiple data centers, such as cloud federations [2], the inter-clouds [3] [4], and hybrid clouds [5] [6], among others. In these approaches, the complexity is revealed when we try to provide resources for a range of users with different needs of applications and services [4], bearing in mind the possibility that the solution to the user request may be in an environment with multiple suppliers with infrastructure managed in completely different forms, i.e., it is a highly heterogeneous computing environment [3] [7].

To tackle problems arising from the allocation of cloud resources and meet the demands of users based on quality of service (QoS) requirements, there is now one of the most discussed topics in cloud computing research field: the intermediation process and task scheduling to cloud computing environments [3].

The recent works which focus their efforts on solving specific problems inherent in cloud environments, such as energy efficient consumption, allocation and migration of virtual machine instances, optimizations in data communication through computer networks within data centers [6] [8] [9] [10], among many other issues, implement, in their methodology, cloud brokers created with strict scheduling policies focused on system balancing for seeking specific goal. However, the new reality of brokering activity for cloud systems is the use of an intermediary architecture represented by a broker that may be multi-objective.

This work relates to the task scheduling and intermediation activity research field, proposing a new Cloud Broker architecture, implemented as simulation entity for CloudSim, working this way as an extension to this cloud computing simulation toolkit. The Broker implemented has the characteristic of openness, i.e., is designed to be coupled to various modes of operation, using as a basis for such implementation the NIST Cloud Computing Reference Model [11] and the operation mode of intermediation services for the experiments.

The remainder of this paper is organized as follows: Section 2 presents the related work reviewed and discussed; Section 3 presents in detail the new Cloud Broker Architecture implemented; Section 4 introduces the design of experiments and the simulation scenario designed to test the Cloud Broker; Section 5 consists of the discussion of the experimental results; Section 6 presents the final conclusion of the work; in Section 7 are presented the acknowledgements and the last section is a list of references.

2. RELATED WORK

The CloudSim Toolkit became an adopted framework for evaluating the test environments of many recent jobs published on the Cloud Computing research field, which mention the tool as relevant and capable of providing the necessary resources for modeling and simulation [12] [13] [14] [15] [16].

In [17], the authors propose a cloud broker architecture for selecting a cloud provider from multiple providers' instances. The cloud broker designed measures the QoS of each provider and sorts them according to the client's request requirements. For differentiation of cloud providers there is the Service Measurement Index (SMI), a relative index calculated to provide the requester a perception gap between the services of different providers. Proper provider selection

technique called TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) is based on the establishment of a ranking for selecting an appropriate cloud provider. The experimental results of this work were obtained from experiments on simulation CloudSim. It conducted a set of experiments considering 6 providers and the authors' conclusion was that the application of the chosen set of techniques allowed an efficient selection of cloud providers based on customer requirements.

In another recent work which deals with the problem of service selection in cloud environments with multiple providers [18], the authors propose a project through a solution approach with a multi agent broker. The Jumper Firefly Algorithm was used in the implementation to reduce the execution time of make span time (response time) through a status table which records past behavior. The validation of all the propositions made at work was carried out with the aid of CloudSim simulation environment. In the experimental results, according to statements of the authors, the Firefly Jumper Mechanism is more effective than the standard Firefly Algorithm and other heuristics that were tested.

In another related work that employ their efforts on rapid and effective execution of jobs sent by users to a cloud computing environment [19], the authors propose a communication framework between the broker elements and the virtual machines (VMs), seeking cost and execution optimal results, that was named Broker Virtual Machine Communication Framework (BVCF). The testing environment was constructed with assistance from CloudSim simulator and its API, creating VM scheduling policies based on cost. In the context of the simulated environment programming were also considered cloudlets scheduling and cloudlets relay, and the review of the implementation of the tasks execution was carried out through the Round Robin and FCFS policies. According to the results obtained in testing and analysis conducted by the authors of work, cost factors and task runtime are always the primary components of the constraints of service quality required by customer requests.

In a job that believes in the growth of the computer market demand and the evolution of the industry into the era of cloud federations and inter-clouds [20], the authors state that the aggregate values to cloud services that will be most valued by customers will be pricing or ticketing policy, the allocation scheme of resources to provide the best performance as the signed service level agreements (SLA). The implementation of the work was carried out with the aid of CloudSim Toolkit version 3.0.3, whereby the authors implemented a broker for cloud federations, which works with the intermediation process, interoperability and negotiation of service requests. According to the authors and the experimental results, it is concluded that the resource allocation model based on QoS and reimbursement worked and successfully demonstrated the applicability and necessity of observation of the QoS degradation in complex environments inter-cloud.

In a work that implements a new scheduling model for cloud computing environments called ICMS (Inter-Cloud Meta-Scheduling) [21], the researchers also created an extension of CloudSim Toolkit which was named SimIC (Inter-Cloud). The goal was to meet the complex simulation scenarios in which inter-clouds contexts are considered and the process of intermediation requests (cloudlets) is done by multiple cloud meta-brokers running dynamic management and real-time workloads received using a standard decision-making to made tasks scheduling. The metrics used for the analysis were Execution Time and RTT (Round Trip Time) and as modification factors of simulated environments were used different user submissions and computational requirements. From the comparative experimental results between the values

returned for the original CloudSim Toolkit and for the SimIC, it was possible to verify and conclude that there were considerable gains in the new algorithms implemented by the ICMS module, especially in the graphs comparing results of execution time metrics.

All related work carried out have important features and contributions related to task scheduling to cloud computing systems using CloudSim. From the observation of all cloud broker implementations made in related work, it is possible to see the existing gap on the issue of standardization of a broker architecture that can be used in order to mix and permit the development and application of various types of scheduling strategies considering multiple service quality factors considered in the related articles. In this paper, the simulation environment includes a QBroker Entity with QoS negotiation for incoming requests, adding a set of desirable characteristics in simulation scenarios that want to provide more realistic and similar results to the real-world cloud systems.

3. CLOUD BROKER PROPOSED ARCHITECTURE

This section will present the cloud broker architecture designed in this work, which was named *QBroker* (QoS Broker). The goal of the implementation was to add features to existing *DatacenterBroker* class in CloudSim API. The version of CloudSim considered in the implementation of the extension was to 3.0.3.

As already mentioned, the main implementation consists of a subclass of *DatacenterBroker* class, which is in *org.cloudbus.cloudsim* package, which was called *QBroker*. It is important to note that *DatacenterBroker* class also has an inheritance relationship with *SimEntity* class belonging to *org.cloudbus.cloudsim.core* package. Through inheritance it was possible to harness and hone, in QBroker class, methods previously inherited from *SimEntity* and *DatacenterBroker* classes.

3.1 QBroker Operation Modes

One of the major new features implemented in the QBroker class is related to the operating modes of this component in cloud architecture. According to the reference model of the NIST [11], the operating modes are the directives that guide how cloud brokers entities must meet customer requests and relate to the resources of service providers. Thus, NIST defines three main models of operation: intermediation, aggregation, and arbitrage. The definition of each of the operation modes of a cloud broker, according to direct reference to NIST [11] model, is presented below:

- **Intermediation:** A Cloud Broker can increase the performance of a given service increasing any specific capacity and providing value-added services to customers. Such performance improvement can be achieved with the management of services, identity management, performance reporting, enhanced security, among others.

- **Aggregation:** A Cloud Broker can combine and integrate multiple services in one or more services. The Broker provides data integration and ensures secure data communication between client and provider.

- **Arbitration**: the arbitration operation mode is similar to the Services Aggregation, with the exception that the services that are grouped are not fixed. In services arbitration, a cloud broker has the flexibility to choose services from multiple providers' services. To perform such activity, for example, the broker can use a credit scoring service to evaluate and select the provider with the best reputation for that type of service requested by the customer.

The new QBroker entity was developed seeking the implementation of all the above operating modes, however, for this specific paper, a version of QBroker is presented in which only the services intermediation operation mode has been developed.

3.2 QBroker Services Intermediation

The process of services intermediation defines some actions for cloud broker in its task as mediator between customers and cloud providers. Increase one or more capabilities of a given service mean improving the quality of service. Therefore, this increase in the providers' service QoS can be achieved in many ways, so that the NIST reference model left open the possibility for the cloud brokers developers.

In this work, the mode of operation of intermediation services was designed to allow that QBroker negotiates the execution of individual requests (cloudlets) with one or more cloud service providers, giving priority to the QoS parameters required by the client and also ensuring the quality of the services, so that, by detecting a degradation of service, the Broker acts allocating new resources (VMs and/or services instances), in order to maintain the satisfactory execution performance and the compliance with other requirements in the requests.

The operating procedure for activity flows related to QBroker Services Intermediation Algorithm is shown in Figure 1, formatted as an UML Activity Diagram (Unified Modeling Language).

Adjustments were made in *Cloudlet* class from *org.cloudbus.cloudsim* package, in which the following class attributes have been added:

- *maxExecutionTime*: variable type *double* in which is stored the maximum execution time or execution deadline.
- *service:* variable type *int* to mark the requested service id.
- *arrivalTime:* variable type *double* that hosts the arrival time of cloudlet at the broker.
- *clientID:* variable type *int* used to identify the source client of a request.
- *sendTime e receiveTime:* are variables of type *double* that are used to store the time of submission of the request by a client and the receipt of cloudlet executed on the client.

It is interesting to notice that this intermediation mode of operation in QBroker is always looking to accomplish the QoS requirement of maximum execution time. This makes the implementation of the operation mode fairly close to the services intermediation definition of NISTCloud Computing Architecture [22].

This characteristic also allows customers to get the results of your requests with quality of service in a hybrid cloud computing environment, always giving priority to the allocation of resources in private cloud and, when needed, allocating resources in the public cloud.

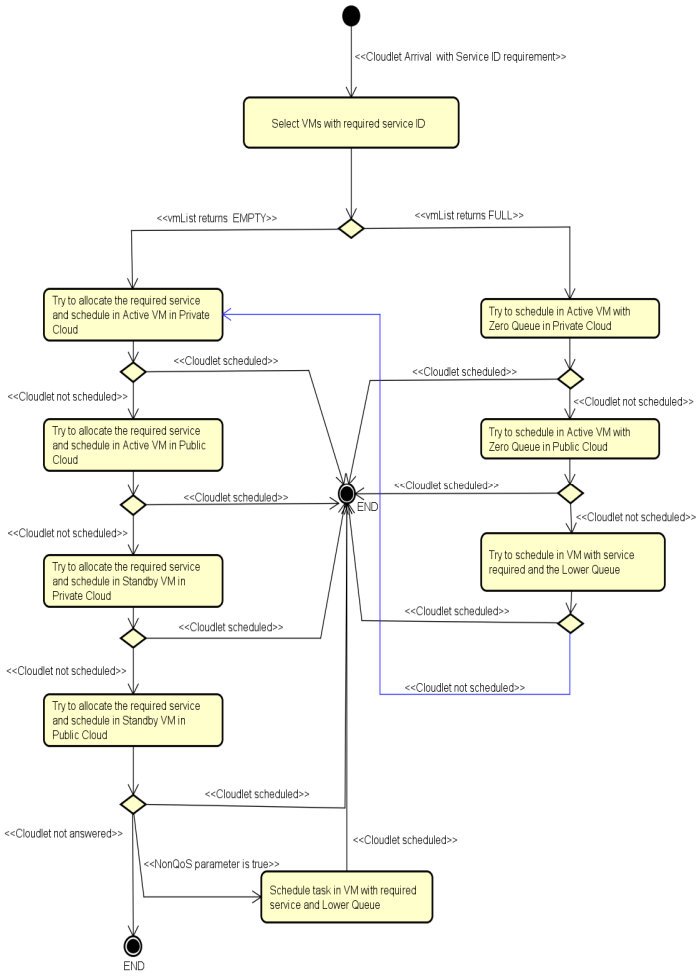


Figure 1: Activity Diagram of QBroker Service Intermediation Operation Mode

3.3 QBroker Class Simulation Events

The simulation entity QBroker has some specific events that were created in addition to support several actions that should occur during the simulation time. For receiving individual requests (cloudlets) the event NEW_CLOUDLET_ARRIVAL was created, through which the cloud broker may receive individual cloudlets during the simulation. It is responsible for receiving task routines, booking and forwarding to the scheduling function and subsequent job submission to a datacenter.

3.4 QBroker class Relationship with other simulation components

To perform its functions during the execution of the simulations, the *QBroker* entity works together with other two important classes implemented in addition: *MetaCloudletScheduler* class and *RequestMonitor* class (which is also an extension of *SimEntity* class). These three classes coexist in the same package named *br.icmc.usp.lasdpc.BeQoS.classes*.

The QBroker class has an instance of RequestMonitor class, in this way, whenever an event of arrival of individual or in group request occurs, the QBroker signals the event of arrival so that the *RequestMonitor* entity receives such notification and account the requests received in cloud broker. The MetaCloudletScheduler class serves as support for QBroker, having all methods that implement the desired scheduling strategies for the cloud computing environment. It is through this class that QBroker is no longer a cloud broker with a rigid systematic task scheduling, offering now the possibility of implementing other scheduling methods. In *MetaCloudlet Scheduler* are methods that allow different types of verification related to resources, whether VMs or services, so that the mediation process is successful.

4. DESIGN OF EXPERIMENTS

In this work were planned three sets of experiments in order to test and demonstrate the features implemented in the intermediation process performed by *QBroker*. All experiments were repeated 10 times, each repetition during 9000 seconds (simulation time based on the clock tick of CloudSim) with 95% confidence interval according to the *T-Student Table*.

4.1 Datacenter and Virtual Machine Configuration

The characterizations adopted for cloud computing simulated scenario were standardized to the three sets of experiments. The scenarios are set up with private cloud or hybrid cloud. The datacenter configurations for private cloud are demonstrated in Table 1.

Table 1: Settings for Private Cloud Infrastructure.

Private Datacenter – Host Configuration	
MIPS/Core:	10000
Cores/Host:	4
RAM:	8000 Mb
Network Bandwidth:	1000 Mbps
Storage:	500000 Mb
OS:	Linux
VMM:	Xen
Total Number of Hosts:	5

The settings of the VMs from private cloud datacenter are shown in Table 2.

Table 1: Settings for Private Cloud VMs.

Private Datacenter – VM Configuration	
MIPS/Core:	10000
PEs Number(Core):	1
RAM:	2000 Mb
Network Bandwidth:	100 Mbps
Image Size:	125000 Mb
VMM:	Xen
Total Number of VMs:	20

The settings used in the simulation scenario with hybrid cloud computing are designed with a public cloud datacenter with settings as demonstrated in Table 3.

Table 3: Settings for Public Cloud Infrastructure.

Public Datacenter – Host Configuration	
MIPS/Core:	20000
Cores/Host:	8
RAM:	32000 Mb
Network Bandwidth:	10000 Mbps
Storage:	1000000 Mb
OS:	Linux
VMM:	Xen
Total Number of Hosts:	2

In the implemented simulation scenario, a total number of 10 VMs on public cloud datacenter was created. The settings for Public VMs are shown in Table 4.

Table 4: Settings for Public Cloud VMs.

Public Datacenter – VM Configuration	
MIPS/Core:	20000
PEs Number(Core):	1
RAM:	4000 Mb
Network Bandwidth:	1000 Mbps
Image Size:	250000 Mb
VMM:	Xen
Total Number of VMs:	10

Also related to cloud computing simulated scenario, the client layer settings were implemented considering a systematic of service demand generation and a fixed amount of customers.

4.2 Service Demand and Client Settings

With regard to service demand generating, a table of service identifiers and their demands in MI (millions of instructions) has been implemented. The service demand for each cloudlet is assigned based on the requested service ID as a specific exponential distribution for each service. The exponential distribution considered has average value of 70000 MI. The total number of possible services, which were considered in the scenario, is 5. It is important to remember that the demand for MI is applied to the length field of each cloudlet, which specifies the size of each task. The services demand values considered in the experiments are listed in Table 5.

Table 2: Service Demand Settings.

Service ID	***Demand (MI)***
S1	30000
S2	50000
S3	70000
S4	90000
S5	110000

The amount of client entities was set to 150 units for all scenarios. The client type configuration, which sets the simulation time client entity operating mode, it was sending requests in real time, meaning that the requests are sent by clients during the course of CloudSim logical clock, creating a more realistic and reliable arrival process to the real world. The generation of service

IDs to be inserted into each request was also made in a random manner considering only 5 services.

To make the heterogeneous service demand, a method in the Client class generates random values that are associated with a service ID as a distribution in percentage. This distribution created can be seen in Table 6.

Table 3: Distribution of services random generation to the requests of client entities

Service ID	*Distribution (%)*
S1	5.0
S2	15.0
S3	60.0
S4	19.0
S5	1.0

Still referring to the configuration of client entities, it is important to note that the QoS attribute considered in each cloudlet was the maximum execution time (*maxExecutionTime*). To obtain the value of QoS constraint field was developed a method in the Client class to ensure that the generation of the maximum execution times are proportional to the size of each cloudlet.

Based on common settings that were explained, it was possible to obtain meaningful simulation results, influencing the response variables considered in the experiments, which will be detailed in the next section,

4.3 Considered Response Variables

For obtaining feedback values in sets of experiments, were selected three response variables that are described below:

- **Response time:** measured in seconds considering the amount of time expended in sending a request to the VM from one provider and its return back to the client.
- **Percentage of Processed Requests**: consider the requests that were processed with *Success* status.
- **Percentage of Unanswered Requests:** consider the requests which could not be met by the cloud broker because not meet the QoS requirement of maximum execution time (*maxExecutionTime*).

5. RESULTS AND DISCUSSION

This section presents information regarding the results of the three sets of executed experimental plans.

5.1 Disclosure of QoS Scenario

The first scenario that will be discussed is the disclosure of QoS. Table 7 summarizes the experimental design created for the scenario in question. Abbreviations found in tables 7, 8 and 9

on the number of VMs field whose acronyms are PRV and PUB, refer, respectively, Private Cloud and Public Cloud.

As can be seen by observing Table 7, experiments with Round Robin Algorithm were compared with experiments using QBroker Services Intermediation Algorithm dealings with or without QoS.

Table 7: Experimental design for disclosure of QoS scenario.

Experiment ID	*Task Scheduling Algorithm*	*Cloud Type*	*Number of VMs*	*Number of Allocated Services*
A	Round Robin	Private	PRV=20	-
B	Round Robin	Hybrid	PRV=20+PUB=10	-
C	Intermediation with QoS	Private	PRV=10	5
D	Intermediation with QoS	Hybrid	PRV=20+PUB=10	5
E	Intermediation without QoS	Private	PRV=20	5
F	Intermediation without QoS	Hybrid	PRV=20+PUB=10	5

In the experiments with intermediation were allocated the five services considered the environment in all VMs in order to make a fair comparison with the Round Robin, which does not have the service selection policy. The results concerning the variable average response time set out in Figure 2.

The obtained results for average response time variable (Figure 2) show that, in private cloud scenarios (experiments A, C and E), QBroker intermediation algorithm proved to be efficient, since in experiment A with Round Robin, the response time was 49.08 seconds while in the experiment E, with intermediation without QoS, obtained better performance with an average time of 45.13 seconds (about 5.8% faster). Still by comparing experiment A with the experiment C, i.e., considering the intermediation with QoS, the performance was even better against the two other experiments, obtaining the value of 18.12 seconds (about 63.08% faster than the experiment A and 59.85% faster than the experiment E).

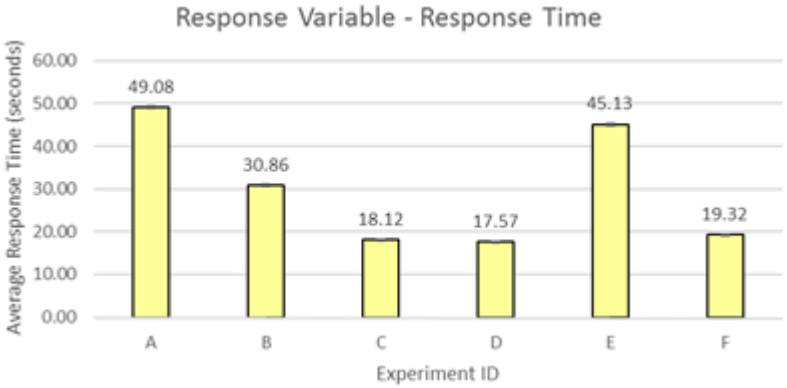


Figure 1: Average response time graph for disclosure of QoS scenario

Although the results with average response time (Figure 2), considering the experiments with hybrid cloud scenarios (experiments B, D and F), the QBroker intermediation algorithm also showed gains in efficiency and performance. The experiment B, which considered using Round Robin had the average response time of 30.86 seconds, while the experiment F considering intermediation without QoS, got 19.32 seconds, which means better performance (about 37.40%

more fast). In experimental examination of experiment D, considering intermediation with QoS, the average value obtained was better than the other two experiments, resulting in 17.57 seconds (about 43.07% faster than Experiment B and 8.9% faster compared to experiment F).

These results corroborate the premise of this paper that the new QBroker Architecture provides performance gains for a major response variables observed by end users of cloud computing systems, that is, the response time for service requests.

It is also possible to visualize differences in how the task scheduling algorithms behave in the simulation scenarios according to the variables of percentage of processed requests and percentage of unanswered requests. According to the results presented by the response variables relating to percentages of processed and missed requests (Figures 3 and 4) stand out from the experiments C and D, which considered scenarios with private and hybrid cloud respectively, using intermediation algorithm with QoS, because it was the only restrictive scenarios on the issue of rejection of requests because of violation of the maximum execution time (*maxExecutionTime*) QoS parameter.

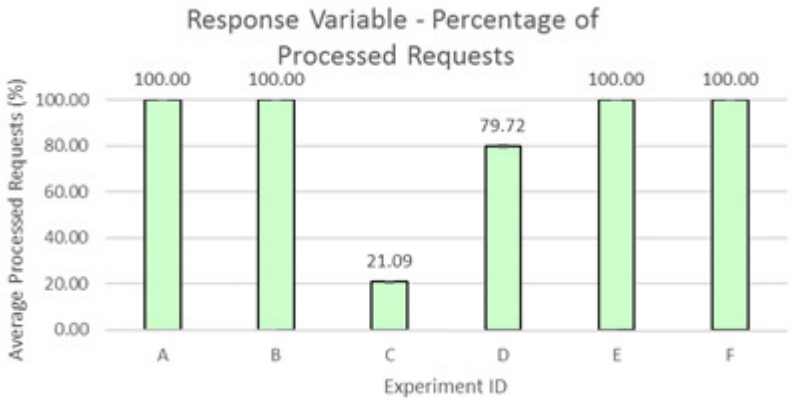


Figure 2: Average percentage of processed requests graph for disclosure of QoS scenario.

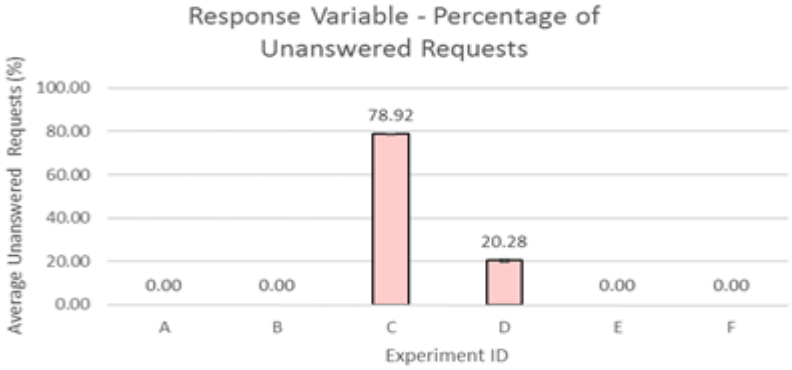


Figure 3: Average percentage of unanswered requests for disclosure of QoS scenario.

In experiment C (according to Figures 3 and 4), as the need arises to keep QoS deadline informed as attribute for each request (cloudlet), QBroker processed 21.09% of the requests sent by clients and rejected others 78.92%. In the experiment D, using the same premise, the QBroker processed 79.72% of the requests and rejected others 20.28%. In other experiments there was no rejection of

requests registered and the response variable percentage of processed requests obtained the constant value of 100%.

5.2 Service Differentiation Scenario

In the second experiment scenario, the objective was to evidence the service differentiation by varying the amount of allocated services in the virtual machines. The characteristic to differentiate services by the use of identifiers approaches QBroker Architecture of cloud brokers compatible with service-oriented architectures (SOA). Table 8 shows the planning of the current scenario of experiments.

According to the experiments plan (Table 8), it is possible to check that the setting of experiments is a variation of the experimental design originally done in disclosure of QoS scenario. The experiments C', D', E' and F' have the same scenario characteristics as, respectively, experiments C, D, E and F, however, the number of services allocated in the machines is different. In the experiments C, D, E and F are allocated 5 services in all instantiated VMs while in experiments C', D', E' and F' the amount of allocated services in the VMs is 2. It should be remembered that in all scenarios where the QBroker used intermediation algorithm, existing services use identifiers numbered from 1 to 5.

Table 8: Design of experiments for service differentiation scenario.

Experiment ID	*Task Scheduling Algorithm*	*Cloud Type*	*Number of VMs*	*Number of Allocated Services*
C	Intermediation with QoS	Private	PRV=20	5
D	Intermediation with QoS	Hybrid	PRV=20+PUB=10	5
E	Intermediation without QoS	Private	PRV=20	5
F	Intermediation without QoS	Hybrid	PRV=20+PUB=10	5
C'	Intermediation with QoS	Private	PRV=20	2
D'	Intermediation with QoS	Hybrid	PRV=20+PUB=10	2
E'	Intermediation without QoS	Private	PRV=20	2
F'	Intermediation without QoS	Hybrid	PRV=20+PUB=10	2

The information of the results of the services differentiation scenario regarding the average response time are shown in Figure 5.

It is possible to see, through the table 8, that the number of services for each VM in this scenario is preset at the beginning of simulation, so there is no occurrence of attempted allocation of new services. In the specific case of the experiments C', D', E' and F', the instantiated services in each VM uses a method of normal distribution for the 5 considered services.

According to Figure 5, for this disclosure of service differentiation scenario, it is possible to note that experiments C and D have the very close results, although not statistically equivalent. Comparing experiments C and C', it can see that C' got an average response time faster with 15.13 seconds. The same situation occurs with the experiments D and D', in which case the experiment D' performed better response time, which value was 14.02 seconds.

Figure 4: Average response time graph for service differentiation scenario.

The justification for these values is precisely the question of the distribution of services, as in the case of experiments C and D, all possible services are instantiated on all scenario's VMs, so that while it may offer more scheduling possibilities for requests, end up making higher the size of average queue, and in this situation, there is a decrease in response time variable and there is also a discard percentage slightly higher.

Experiments with only two services and use of intermediation with QoS (as Figure 5), i.e., C' and D', although become the most restrictive scenario for scheduling options of requests for VMs, generate an average queue time differentiated of a VM to another, because those services whose demand exponential function are larger are not instantiated on all VMs, leading to this situation in particular, a better performance in response time variable.

Also relating to information from experiments in Figure 5, in experiments E, F, E' and F', the results have another positioning. As in experiments E and F has all instantiated services in all VMs of the scenarios and the availability ends thus being wider, and, as already explained, considering that the last activity of intermediation without QoS is schedule the request to the VM that has the service requested instantiated with the lower queue, in such cases, scenarios with more services offer more scheduling opportunities, which makes the values of average times of E and F the experiments, i.e., 45.13 seconds and 19.32 seconds respectively, perform better than the experiments E' and F' having two instantiated services in all scenario's VMs.

The figures 6 and 7 have the performance graphs of percentage of processed and missed requests to the current experiments scenario.

To disclosure a little more the argumentation for the average response time variable, it is possible to observe, as figures 6 and 7, that experiments C and D gave a lower value in terms of processed requests and in turn, higher percentage of unanswered requests (figure 7) as arguments already provided on considerations involving the response time variable.

According to figures 6 and 7, in other experiments (E, F, E' and F') which do not consider the QoS parameter *maxExecutionTime*, always get 100.0% of processed requests, so that there are no unanswered requests.

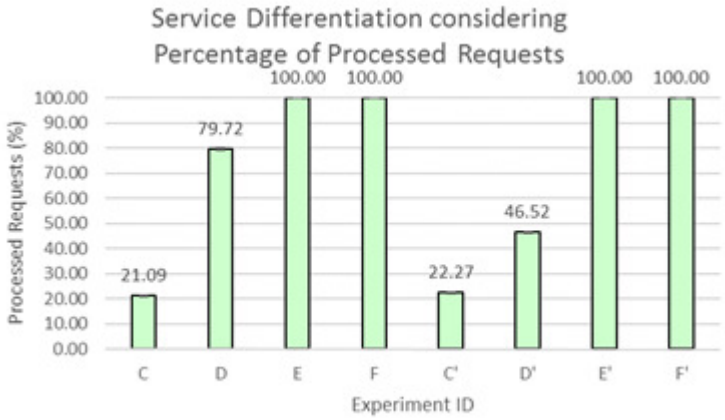


Figure 6: Average percentage of processed requests graph for service differentiation scenario

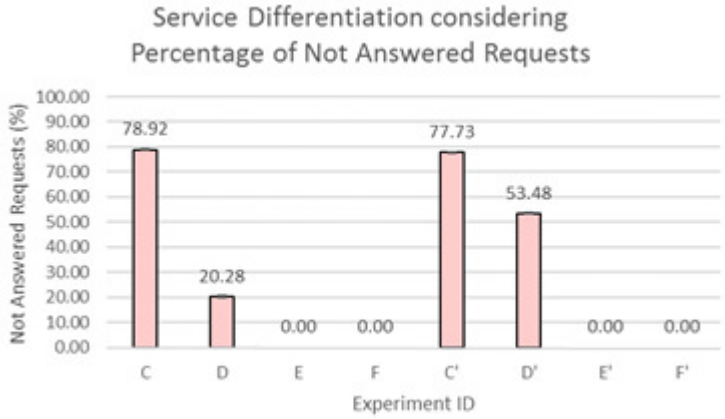


Figure 7: Average percentage of unanswered requests graph for service differentiation scenario

5.3 Dynamic Service Allocation Scenario

In the third experiments scenario the main objective was to highlight the dynamic allocation of services at runtime conducted by QBroker. Table 9 has the experiments planning information created for the experimental scenario explained.

Table 9: Experimental design of dynamic services allocation scenario.

Experiment ID	*Task Scheduling Algorithm*	*Cloud Type*	*Number of VMs*	*Number of Allocated Services*
C'	Intermediation with QoS	Private	PRV=20	2
D'	Intermediation with QoS	Hybrid	PRV=20+PUB=10	2
E'	Intermediation without QoS	Private	PRV=20	2
F'	Intermediation without QoS	Hybrid	PRV=20+PUB=10	2
C''	Intermediation with QoS	Private	PRV=(10 + 10 Stdby)	2
D''	Intermediation with QoS	Hybrid	PRV=(10 + 10 Stdby) + PUB=(10+10 Stdby)	2
E''	Intermediation without QoS	Private	PRV=(10 + 10 Stdby)	2
F''	Intermediation without QoS	Hybrid	PRV=(10 + 10 Stdby) + PUB=(10+10 Stdby)	2

According to the experiments plan (Table 9), it is possible to note the fact that were made a combination of experiments with fixed number of services (C', D', E' and F') with four other experiments that perform dynamic allocation of services. It can also to note that in the private cloud experiments, only 5 VMs have 2 instantiated services while the other 15 VMs remain in standby state. In the scenario with hybrid cloud, private cloud is initialized with the same previous configuration and the public cloud is initialized with all the VMs in standby state.

The results concerning the average response time variable for current scenario are shown in Figure 8.

From graph analysis, it can be observed that the experiments which consider intermediation algorithm with QoS (C', D', C'' and D'') have a difference in performance, is noted that the experiments with dynamic service allocation the response time was longer.

The response time in experiment C', which considered static service allocation and private cloud was 14.66% faster than C'', with dynamic service allocation. A similar situation occurs between experiments with hybrid cloud in the scenarios, i.e., the experiment D', considering static service allocation, obtained response time of 17.67% faster than the experiment D'', which used dynamic service allocation. This result was expected because, at the beginning of the execution of simulation experiments, the experiments C'' and D'' has only 5 VMs available for task scheduling, so the dynamic allocation of services is executed when there is real necessity due to the breach of QoS parameter maximum execution time.

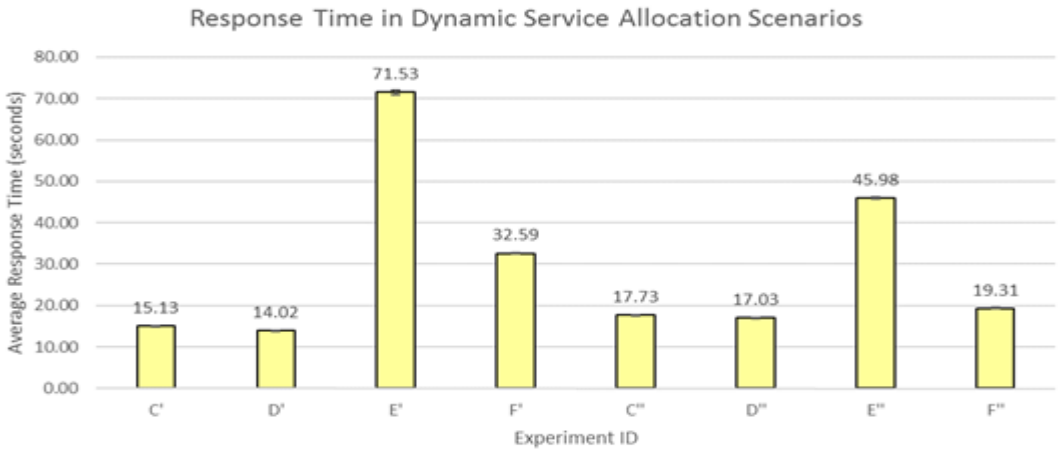


Figure 5: Average response time graph for dynamic service allocation scenario.

Still referring to Figure 8, the experiments that have been configured with intermediation without QoS (E', F', E'' and F'') have a different result because, in this particular case, the experiments with dynamic allocation of services have outstanding difference, with better performance. Experiments E' and F' start with 2 services using a normal distribution. Due to this justified reason, the experiments E 'and F' end up having a lower performance for response time variable because the arrangement of services is predefined at the start of the simulation.

The experiments E'' and F'', have only 5 VMs that are initially initialized with services using the same uniform distribution method. Thus, by effecting on demand service allocation, they have

significant advantage, since the services are allocated on the basis of real need and as services are required in requests.

Figures 9 and 10 present the result of information of variable percentage of processed and missed requests. The experiments in which used the intermediation algorithm without QoS (E', F', E" and F") have a similar behavior, i.e., the variable percentage of processed requests in these experiments was 100.0% and there was no unanswered request.

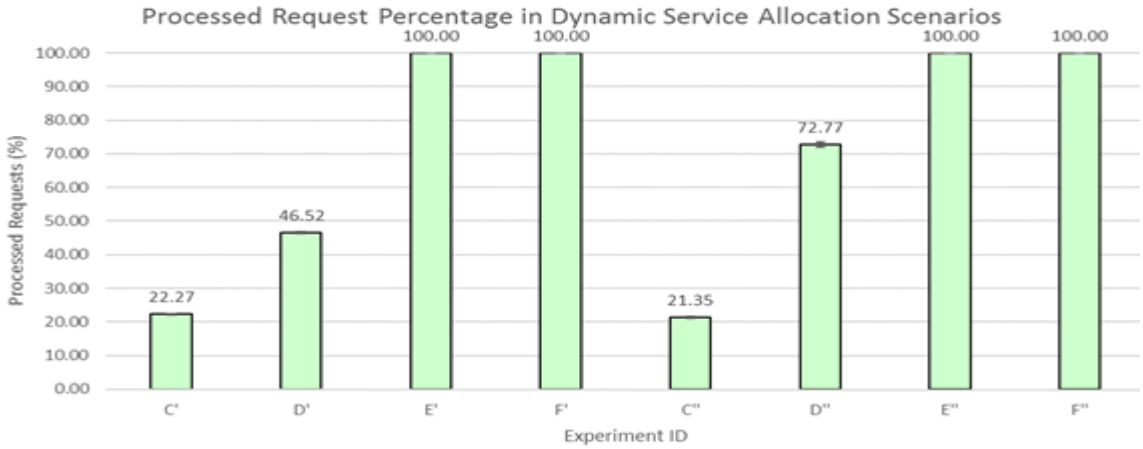


Figure 6: Average percentage of processed requests graph for dynamic service allocation scenario.

Already in the experiments with intermediation with QoS, in the case of experiments considering private cloud C' and C", their values have percentages of processed and unanswered requests next, revealing a similar behavior in the restricted environment of private cloud resources. As for experiments D' and D", which consider hybrid cloud, the experiment D" achieved a better result because, processed a higher percentage of requests, this takes place, as already explained, because of the dynamic service allocation at runtime, what revealed a QBroker feature, that makes the attendance to virtual clients more profitable and causes almost an adaptive effect when you look at the records of the allocation of services performed during the execution of the experiment in CloudSim output report.

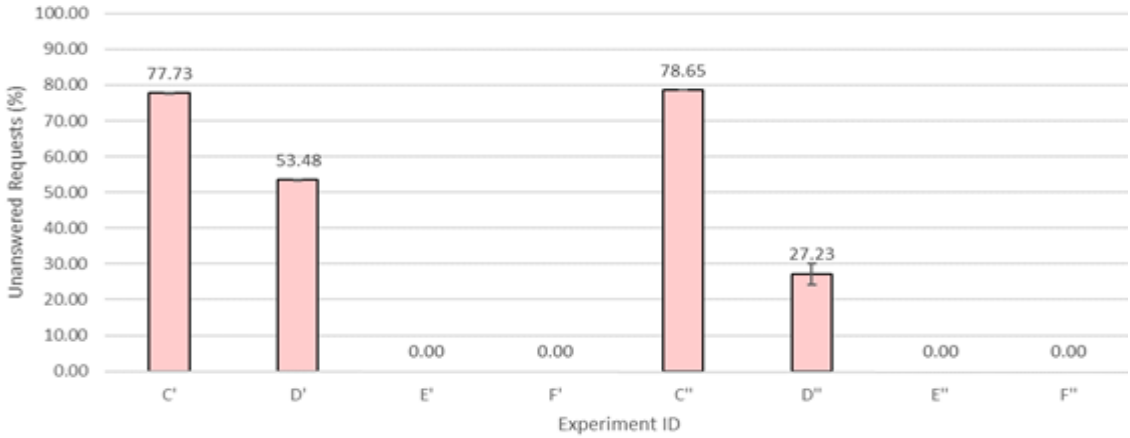


Figure 7: Average Percentage of unanswered requests for dynamic service allocation scenario.

The results of this scenario show that the resource of dynamic service allocation, present in QBroker service intermediation operation mode, is an important differential in the reproduction of real situations of task scheduling to cloud computing systems.

6. CONCLUSION

In this work was presented a cloud broker architecture that provides several features to obtain QoS in hybrid cloud computing environments. To this end, it was implemented, based on the service intermediation definition of NIST cloud computing reference model, a task scheduling policy that considers maximum deadlines for execution of service requests, the allocation control and management of the amount number of services in each VM and the dynamic service allocation on-demand during the execution of simulations. These three key features help the intermediator component of the architecture, that is, help the QBroker to increase the QoS of the services requested on demand, a fact that has been proven through design of experiments performed and presented in three scenarios.

It is worth noting that the Broker is a component that is part of a cloud computing architecture called CloudSim BEQoS (Bursting Energy and Quality of Service), developed by the Laboratory of Distributed Systems and Concurrent Programming (LaSDPC), which is linked to the ICMC University of São Paulo Campus of São Carlos. The results presented in this work highlight the functionality of QBroker operation mode named as service intermediation (with or without QoS). As the information presented from experimental results, it is possible to see the interesting contributions on the simulation of hybrid cloud computing environments through CloudSim coupled to QBroker, MetaCloudletScheduler and other components of BEQoS Architecture.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Brazilian Foundations FAPESP, CNPq and CAPES for the projects under development at the Distributed System and Concurrent Program Group of the Computes Systems Department at ICMC - USP.

REFERENCES

[1] S. Moore, “Gartner Says Worldwide Cloud Infrastructure-as-a-Service Spending to Grow 32.8 Percent in 2015,” 2015. [Online]. Available: http://www.gartner.com/newsroom/id/3055225. [Accessed: 07-Oct-2015].

[2] M. Salama and A. Shawish, “A QoS-Oriented Inter-cloud Federation Framework,” 2014 IEEE 38th Annu. Comput. Softw. Appl. Conf., no. Cc, pp. 642–643, Jul. 2014.

[3] S. Sotiriadis, N. Bessis, and N. Antonopoulos, “Towards Inter-cloud Schedulers: A Survey of Meta-scheduling Approaches,” 2011 Int. Conf. P2P, Parallel, Grid, Cloud Internet Comput., pp. 59–66, Oct. 2011.

[4] M. Aazam and E. N. Huh, “Inter-cloud Media Storage and Media Cloud Architecture for Inter-cloud Communication,” 2014 IEEE 7th Int. Conf. Cloud Comput., pp. 982–985, Jun. 2014.

[5] M. H. Sqalli, M. Al-saeedi, F. Binbeshr, and M. Siddiqui, “UCloud: A simulated Hybrid Cloud for a university environment,” 2012 IEEE 1st Int. Conf. Cloud Netw., pp. 170–172, Nov. 2012.

[6] V. Bagwaiya and S. K. Raghuwanshi, "Hybrid approach using throttled and ESCE load balancing algorithms in cloud computing," 2014 Int. Conf. Green Comput. Commun. Electr. Eng., pp. 1–6, Mar. 2014.

[7] M. Aazam and E.-N. Huh, "Broker as a Service (BaaS) Pricing and Resource Estimation Model," 2014 IEEE 6th Int. Conf. Cloud Comput. Technol. Sci., pp. 463–468, Dec. 2014.

[8] M. Nir, A. Matrawy, and M. St-Hilaire, "An energy optimizing scheduler for mobile cloud computing environments," 2014 IEEE Conf. Comput. Commun. Work. (INFOCOM WKSHPS), pp. 404–409, Apr. 2014.

[9] R. S. Moorthy, T. S. Somasundaram, and K. Govindarajan, "Failure-aware resource provisioning mechanism in cloud infrastructure," 2014 IEEE Glob. Humanit. Technol. Conf. - South Asia Satell., pp. 255–260, Sep. 2014.

[10] E. Hwang and K. H. Kim, "Minimizing Cost of Virtual Machines for Deadline-Constrained MapReduce Applications in the Cloud," 2012 ACM/IEEE 13th Int. Conf. Grid Comput., pp. 130–138, Sep. 2012.

[11] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, "NIST Cloud Computing Reference Architecture Recommendations of the National Institute of Standards and."

[12] R. N. Calheiros, R. Ranjan, A. Beloglazov, and A. F. De Rose, "CloudSim : a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," no. August 2010, pp. 23–50, 2011.

[13] B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "CloudAnalyst: A CloudSim-Based Visual Modeller for Analysing Cloud Computing Environments and Applications," 2010 24th IEEE Int. Conf. Adv. Inf. Netw. Appl., pp. 446–452, 2010.

[14] X. Li, X. Jiang, P. Huang, and K. Ye, "DARTCSIM : AN ENHANCED USER-FRIENDLY CLOUD SIMULATION SYSTEM BASED ON CLOUDSIM WITH," 1857.

[15] S. Long and Y. Zhao, "A Toolkit for Modeling and Simulating Cloud Data Storage: An Extension to CloudSim," 2012 Int. Conf. Control Eng. Commun. Technol., pp. 597–600, Dec. 2012.

[16] S.-M. Jung, N.-U. Kim, and T.-M. Chung, "Applying Scheduling Algorithms with QoS in the Cloud Computing," 2013 Int. Conf. Inf. Sci. Appl., pp. 1–2, Jun. 2013.

[17] R. Achar and P. S. Thilagam, "A broker based approach for cloud provider selection," 2014 Int. Conf. Adv. Comput. Commun. Informatics, pp. 1252–1257, Sep. 2014.

[18] N. G. and J. G., "A Multi-agent Brokering Approach and Jumper Firefly Algorithm for Job Scheduling in Cloud Computing," 2014 Int. Conf. Intell. Comput. Appl., pp. 52–58, Mar. 2014.

[19] G. Raj and S. Setia, "Effective Cost Mechanism for Cloudlet Retransmission and Prioritized VM Scheduling Mechanism over Broker Virtual Machine Communication Framework," vol. 2, no. 3, pp. 41–50, 2012.

[20] M. Aazam and S. Korea, "Advance Resource Reservation and QoS Based Refunding in Cloud Federation," pp. 139–143, 2014.

[21] S. Sotiriadis, N. Bessis, and N. Antonopoulos, "Towards Inter-cloud Simulation Performance Analysis: Exploring Service-Oriented Benchmarks of Clouds in SimIC," 2013 27th Int. Conf. Adv. Inf. Netw. Appl. Work., pp. 765–771, Mar. 2013.

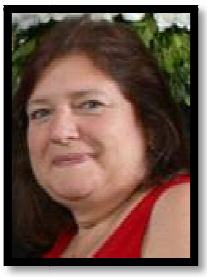
[22] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST cloud computing reference architecture," Proc. - 2011 IEEE World Congr. Serv. Serv. 2011, pp. 594–596, 2011.

AUTHORS

Mário Henrique de Souza Pardo received bachelor's degree in Systems Analysis from the University of the Sacred Heart (USC), Bauru / SP, Brazil, in 2001. concluded a master's degree in Computer Science from UNIVEM, Marilia / SP, Brazil in 2006. He is currently a PhD student in research line of Distributed Systems and Concurrent Programming from the University of São Paulo (USP), São Carlos / SP, Brazil, supervised by Professor Dr. Regina H. C. Santana. His current research interest is focused on the study of Cloud Computing, specifically for task scheduling with QoS for complex cloud computing environments. Also deals with computer systems performance evaluation and simulation of cloud computing through discrete event simulation.

Adriana Molina Centurion received the BS degree in Computer Science from Marilia University (UNIMAR) in 1995; the MS degree (in 1998) and PhD degree (in 2015) in Computer Science and Computational Mathematics from the University of Sao Paulo (USP). She has experience in Computer Science with emphasis in Distributed Computer Systems and Performance Evaluation and 10 years of experience in management of projects and services in the area of Information Technology. She currently is professor at Institute of Education, Science and Technology of Sao Paulo (IFSP). Her research interests include Performance Evaluation, Distributed Systems, Service Oriented Architecture, Cloud Computing, Simulation, Workload Modeling and Burstiness Phenomenon.

Paulo Sérgio Franco Eustáquio graduated Bachelor in Computer Science from the Pontifical Catholic University of Minas Gerais (PUC), Pocos de Caldas / MG, Brazil, holds a Master degree in Computer Science from the University of São Paulo (USP), São Carlos / SP, Brazil. It is a PhD student at Institute of Mathematics and Computer Sciences (ICMC) at USP of São Carlos / SP, Brazil, supervised by Professor Dr. Sarita M. Bruschi. His research interests are: Cluster, Grid and Cloud Computing, Web Servers Architecture, intake systems and task scheduling for distributed computing systems, client-side QoS and provider-side QoS considering Green Computing and efficient energy consumption for Computing Cloud environments.



Regina Helena Carlucci Santana graduated in Electrical Electronic Engineering from the School of Engineering of São Carlos (1980), Master degree in Computer Science from the Institute of Mathematical Sciences of São Carlos (1985) and PhD in Electronics and Computing - University of Southampton (1989). She is currently Associate Professor at the University of São Paulo. She has expertise in Computer Science, with emphasis on Performance Evaluation, acting on the following topics: performance measurement, simulation, distributed simulation, tasks and process scheduling and parallel computing. Other topic of her interest in research is Distributed Computational Systems Architecture involving Cluster, Grid, Cloud Computing and others.

Sarita Mazzini Bruschi graduated in Bachelor of Computer Science from Paulista State University "Julio de Mesquita Filho" (1994), Master degree in Computer Science from the University of São Paulo (1997) and PhD in Computer Science from the University of São Paulo (2002). She is currently Doctor Professor MS3 RDIDP at the University of São Paulo. She has expertise in Computer Science, with emphasis on Performance Evaluation, acting on the following topics: performance evaluation, simulation, tasks and process scheduling in Cloud Computing, Green Computing, Educational Environments and Operating System.

Marcos José Santana graduated in Electrical Electronic Engineering from the School of Engineering of São Carlos (1980), Master degree in Computer Science from the Institute of Mathematical Sciences of São Carlos (1985) and PhD in Electronics and Computing - University of Southampton (1989). He is currently Associate Professor at the University of São Paulo. He has expertise in Computer Science, with emphasis on performance evaluation, acting on the following topics: performance evaluation, web services, cluster computing, grid computing, cloud computing, process scheduling, parallel computing, simulation and load balancing for distributed systems. Coordinator of Computer Engineering at ICMC since 2002 to 2011 and Chief of the Computer Systems Department since 2010.

DESIGN AND IMPLEMENT A NEW CLOUD SECURITY METHOD BASED ON MULTI CLOUDS ON OPEN STACK PLATFORM

Mohamad Reza Khayyambashi and Sayed Mohammad Hossein Mirshahjafari and EhsanShahrokhi

Department of Computer, Faculty of Engineering,
University of Isfahan, Isfahan-Iran
M.R.Khayyambashi@eng.ui.ac.ir, m.mirshah@irisaco.com,
ehsantux@gmail.com

ABSTRACT

Deployment of using cloud services as a new approach to keep people's platforms, Infrastructure and applications has become an important issue in the world of communications technology. This is a very useful paradigm for humans to obtain their essential needs simpler, faster ,more flexible, and safer than before. But there are many concerns about this system challenge. Security is the most important challenge for cloud systems. In this paper we design and explain the procedure of implementation of a new method for cloud services based on multi clouds on our platform which supplies security and privacy more than other clouds. We introduce some confidentiality and security methods in each layer to have a secure access to requirements. The architecture of our method and the implementation of method on our selected platform for each layer are introduced in this paper.

KEYWORDS

Cloud Computing, Security, MultiClouds, Secure Cloud Architecture, Public Cloud, Personal Cloud

1. INTRODUCTION

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." This is NIST's (National Institute of Standards and Technology) definition of cloud computing. Definition of cloud computing is based on five attributes: multi tenancy (shared resources), massive scalability, elasticity, pay as you go, and self-provisioning of resources.Cloud data can take many forms. For example, for cloud-based application development, it includes the application programs, scripts, and configuration settings, along with the development tools. For deployed applications, it includes records and other content created or used by the applications, as well as account information about the users of the applications.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 83–91, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60208

Data that is stored on cloud must be secured while at rest, in transit, and inuse, and access to the data needs to be controlled. Standards for communications protocols and public key certificates allow data transfers to be protected using cryptography.Currently, the responsibility for cryptographic key management falls mainly on the cloud service subscriber.

Threewidely referenced service models have evolved:

- Software-as-a-Service (SaaS) enables a software deployment model in which one or more applications and the computing resources that run them are provided for use on demand as a turnkey service. It can reduce the total cost of hardware and software development, maintenance, and operations.

- Platform-as-a-Service (PaaS) enables a software deployment model in which the computing platform is provided as an on-demand service which applications can be developed upon and deployed. It can reduce the cost and complexity of buying, housing, and the managing of hardware and software components of the platform.

- Infrastructure-as-a-Service (IaaS) enables a software deployment model in which the basic computing infrastructure of servers, software, and network equipment is provided as an on-demand service upon which a platform to develop and execute applications can be founded. It can be used to avoid buying, housing, and managing the basic hardware and software infrastructure components.

In September2009, IDC Enterprise Panel held its annual survey on cloud computing organizations about the most important challenges of cloud services. The result of this survey showed security among people who want to use cloud services is the most important challenge. Figure 1 shows the result in percentage of the survey's concerns.

So if we want cloud computing as a useful service we should provide confidentiality and security for it to reduce this concern. Otherwise clouds can't reach a good position among people for using.

Cloud users and providers have many concerns about using it as a new technology. When considering using a cloud service, the user must be aware of the fact that all data given to the cloud provider leaves his/her own control and protection sphere.

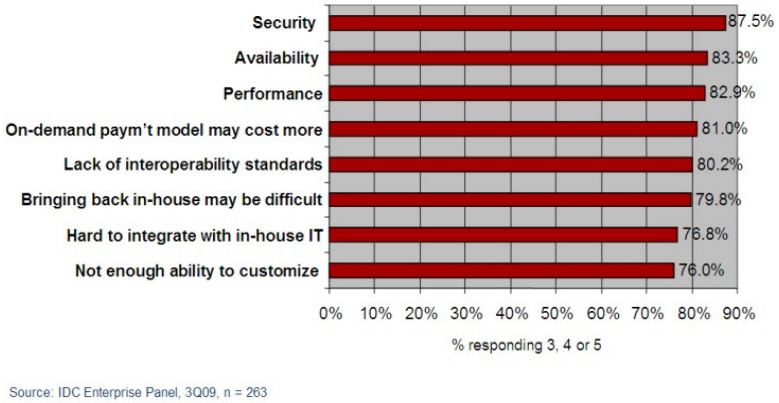


Figure 1. Result of percentage of survey's concerns by IDC, 2009

Even more so, if deploying data-processing applications to the cloud (via IaaS or PaaS),a cloud provider gains full control on these processes. If an attacker is able to intrude the cloud system, all of the data and processes of users operating on that cloud system, may become subject to malicious actions by that attacker. So the methods that cloud providers use to protect their clouds from threats and also the policy for accessing to the cloud by the users must be declared.

2. SECURITY THREATS FOR CLOUDS

As described security is the most important concern in cloud computing. This issue is organized into several general categories: trust, architecture, identity management, software isolation, data protection, and availability. So many threats to cloud computing can exist.

CSA(Cloud Security Alliance) is a research group on cloud security. They released their research results as “Top Threats to Cloud Computing” in 2010 in which they introduce the 7 top threats to clouds security challenges. The top threats they released consist of :

- Abuse and Nefarious Use of Cloud Computing
- Insecure Application Programming Interfaces
- Malicious Insiders
- Shared Technology Vulnerabilities
- Data Loss/Leakage
- Account, Service & Traffic Hijacking
- Unknown Risk Profile

The purpose of these are to provide desirable context to assist organizations in making educated risk management decisions regarding their cloud adoption strategies.

For the common case of a cloud provider hosting and processing all of its user’s data, an intrusion would immediately affect all security requirements: accessibility, integrity, and confidentiality of data and processes may become violated,and further malicious actions may be performed on behalf of the cloud user’s identity.

So providing a strong trusting relationship between the cloud providers and the cloud users is still indispensable.Providers should make and represent their security solutions for cloud threats to decrease consumer and organization's concerns.

Security must be provided in each layers of clouds. If we only have a safe physical layer, users will still have concerns about network layers, application layers and others. Although security approach should being applied on all cloud services contains infrastructure-as-a-service (IaaS) security, providers’ platform-as-a-service (PaaS) security and software-as-a-service (SaaS) security.

So our method should be complete and shouldn’t allow any attackers to access or change our cloud's content.

Security problems for clouds do not have any real comprehensive solutions and existing cloud security is in its infancy. There is a need for an approach to cloud security that is holistic, adaptable, and reflects client requirements.

Cloud providers and researchers all over the world worked on this issue and tried many solutions to reduce security risks of the cloud and they reached some solutions for each threat such as authentication, authorization and identification to provide confidentiality, isolation and encryption of cloud data in other layer. But cloud computing becomes bigger and bigger and its challenges grow too.

3. SECURE CLOUD BASED ON MULTICLOUDS METHODOLOGY

Cloud costumers and users worry about using this phenomenon today. We decided to suggest a useful method to decrease cloud's security threats of which we then designed its architecture. And last,we used a platform to implement our security model . We will now explain these steps.

Our method is based on multiple clouds. In other words we use this model to create a secure cloud. We think this model increases our cloud's transparency for consumers and decreases some user's concern about the complexity of clouds and their type of needs of our requests for variety of access level. We have some clouds in our model's architecture that user's data has been put on them. Our clouds are nested and each of them have an access level that according to the needs, this data put on each of them.

For choosing which cloud layer we want to put our data in, first after connecting to the server it asks us about which cloud we want to save our data. In other word we design a contract that forces clients to choose their level of storage and give their username and convert it to hash and save it. The server should sign an international security communication protocol mutuall to ensure user data security and save or recover their data in any circumstances. As we described one of the most common compliance issues facing an organization is data location. In our method we use external audits and security certifications to alleviate this concern. These certifications are different in various countries and it depends on where our method swere used for example DSS(Data Security Standards), The EC Data Protection Directive, GLBA (The Gramm-Leach Bliley Act), CPNI (The FCC Customer Proprietary Network Information rules) and so on.

Availability is one of our main targets for our secure cloud method. Availability means that an organization has its full set of computing resources accessible and usable at all times. It can be affected temporarily or permanently, and a loss can bepartial or complete. Denial of service attacks,equipment outages, and natural disasters are all threatsto availability.The level of reliability of our cloud service and also its capabilities forbackup and recovery is taken into account inthe organization’s contingency planning to address the restoration and recovery of disrupted cloud layers and operations, using alternate services, equipment,and locations.

In our method we describe a cloud that is in the outer surface. We named this cloud "Cloud by public access" and called it CBPA as abbreviation. This is a public cloud. All of our clouds are in this. Data and application that put in CBPA don’t have any protection. So in this layer of our cloud, typically, we have some costumer's data, open source programs and applications and platforms which they don’t want to do any security method or authentication on it. (So developers don't put any preventive method from intruders attack on it. Here is a diagram of our cloud in which CBPA is determined.

Notice that everyone can have access to all things that are put in this layer so all of the data that’s put in this layer is not secure and costumers shouldn't put their important data on it. This is

appropriate for only open source applications or infrastructures or data that they want to show to all costumers. This access level can increase transparency of our cloud and access to this layer is faster than other layers but it has less security than other layers of our method.

Besides authentication, the capability to adapt user privileges and maintain control over access to resources is also required, as part of identity management. Standards like the Extensible Access Control Markup Language (XACML) can be employed to control access to cloud resources, instead of using a service provider's proprietary interface. XACML focuses on the mechanism for arriving at authorization decisions, which complements SAML's focus on the means for transferring authentication and authorization decisions between cooperating entities. Messages transmitted between XACML entities are susceptible to attack by malicious third parties, making it important to have safeguards in place to protect decision requests and authorization decisions from possible attacks, including unauthorized disclosure, replay, deletion and modification.

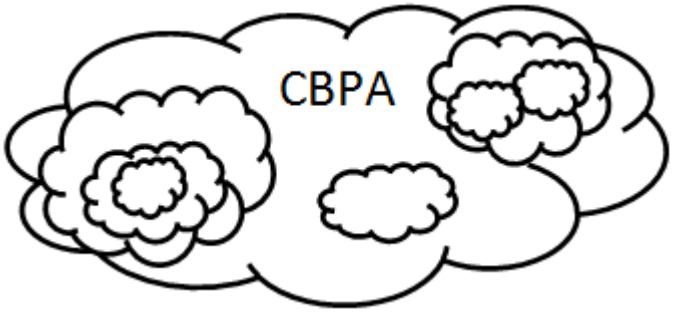


Figure 2. Secure Cloud Architecture base on Multi Clouds

This cloud type includes two types of private clouds: "cloud by group access (CBGA)" and "cloud by personal access (CBPeA)"that_are in the CBPA.

Another cloud in our model that we want to define is "cloud by group access (CBGA)" that located in CBPA. In other wordsthis layer is a branch of our multi cloud model that is in cloud by public access and provides different access level for data.In this layer we considered some security solutions for accessing the contents.

Group access means having some users in a group by identical access level. This model is useful for companies, organizations or any groups that want to have a cloud to put their data in platforms on it for their clients to read, write and edit their information. In our design for this cloud we put some security proceeding to have a more secure level. As we explained before for access to secure clouds we should provide confidentiality. So in this level we supplyconfidentiality by three security methods: Identification, authentication and authorization and supply cloud security by isolation of data. As a service provider we have to ensure dynamic flexible delivery of service and isolation of user resources. For doing this security level we used OpenStack platform and it used two layers for isolating data.

This method here is performed in two levels: first we do these work to authenticate the user that was in this CBGA which this level eliminate one of the most important concerns of cloud consumers but after this security level because we want attackers or Intruders can't access to group's information or to prevent information access by illegal clients, when one of our privileged

clients loses his/her public keys we introduce a second level for this type of cloud that is used to authenticate person who is in the group. This authentication method is used for group members to secure their access on groups and make group safe.

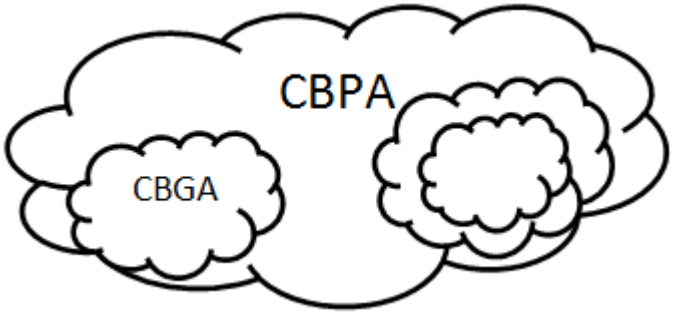


Figure 3. Architecture of CBPeA and ABGA

Another cloud _ we designed is "cloud by personal access (CBPeA)" that is suitable for saving personal data.

In this cloud we have some solutions to keep data secure too. Usage of this type of cloud is more than other types because all of the consumers can save their information on this cloud layer which only they can access and it provides confidentiality and isolation of data like CBPA. But we have some difference in this cloud designing. CBPeAconsists another cloud in itself named "cloud by secure personal access (CBSPeA) that is more secure than normal personal access. In this type we designed encryptionfor data that consumers want to save in addition to the authentication, authorization, identification and isolation.

So we have a secure cloud in this layer that no one can access _ unless main users whose data it is. This cloud is appropriate for user information that is personal and they want to be more secure than other information for example they can put their confidential documents, personal tools or anything that they don't want anyone to access _. Here is the view of this cloud type in our model.

4. IMPELEMENTATION OF METHOD ON OPENSTACK PLATFORM

So we designed our method and explained our architecture.For implementing our cloud model we use OpenStack platform. OpenStack offers open source software to build public and private clouds.This platform has three main components: Compute, Object Storage, and Image Service. OpenStack Compute is a cloud fabric controller, used to start up virtual instances for either a user or a group. It's also used to configure networking for each instance or project that contains multiple instances for a particular project. OpenStack Object Storage is a system to store objects in a massively scalable large capacity system with built-in redundancy and failover. OpenStack Image Service is a lookup and retrieval system for virtual machine images. Our public and private clouds have these components. The OpenStack Compute component of our public cloud can control & manage the inner private clouds. It connects to the compute component of the private clouds. The following diagram shows the basic relationships between the projects, how they relate to each other:

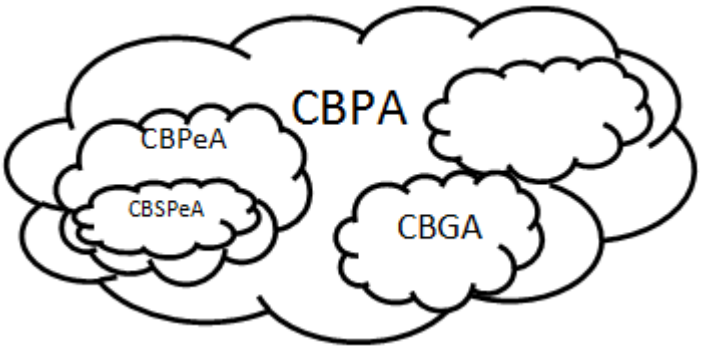


Figure 4. Cloud Secure Architecture with cloud layers names

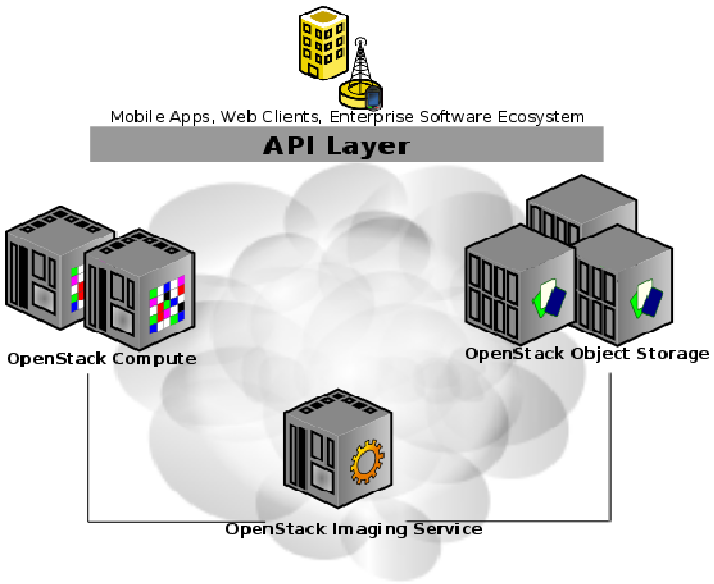


Figure 5. API Layer on openStack platform

In OpenStack compute component we have three subcomponents: Cloud controller, Cluster controller, and Node Controller. The main task of Cloud Controller(CC) are management & controlling the current cloud & the other clouds which are connected to main cloud. This component have a relation to other components. They are Cluster Controller(CLC),Object Storage, and Image Service. These relations are done by REST/SOAP messaging over http protocol. Cluster Controller(CLC) is the manager of the clusters. A cluster is a collection of computers(Nodes) which have been connected to a main server(Frontend). In a cloud we could have one or some clusters. Object Storage has a server that manage the space of the storage of our cloud, we name this server Storage Controller(SC). Image Service has a server for managing the instances of virtual machines and saving of images, we name this server Instance Controller(IC).Each of these server applications run as a daemon (A computer program runs as a background process) in a Linux base OS. Since the each cloud computing service needs a graphical user interface web application for accessing to it, we need a web server for saving & running the web application scripts(We use PHP). This web server is usually in CC server, but it could be in the other assigned server or an external server(Host).This web based interface has a relation to the CC server and uses the primary authentications for accessing to it. In our model,

the main cloud which is public(CBPA) has a CC server that has a connection to its CLCs,SCs & ICs. We assigned for each server a static class C IP(eg. 192.168.100.1 for CC,192.168.100.2 for CLC,192.168.100.3 for SC & 192.168.100.4 for IC). The inner clouds which are private(CBGA,CBPeA,CBSpeA) have these components too. The CC of the main cloud has connection to the CCs of these clouds. In fact one of the tasks of our main CC is management of the inner clouds CC. The procedure is that the user enters his/her username & password in web application UI and after a authentication He/She can se the cloud. In this mode the user can use the public services in cloud such as a application programs(SaaS) ,Platforms(Paas) and a resources(IaaS).If the user(Often a organization) want to has a private cloud, they can use the inner private clouds. For accessing to these they are authenticated again. Each of the authentication actions are done via the components of Object Compute(CC). The users of each group or organization have access to their clouds by group access(CBGA) data jointly. For accessing to each data we define a policy for each of them. It means that which user or group can access to that data or instance. This is what we name it Authorization in security. These authorization are done via the components of Object Storage(SC) and Image Service(IC). The isolation of the data is done by these components too.

5. CONCLUSION

Cloud computing will soon be a big approach in the entire world that conquers all ancient technology. But it depends on removing all concern about this challenge. The migration to a cloud computing environment isin many ways an exercise in risk management. Bothqualitative and quantitative factors apply in ananalysis. An appropriate balancebetween the strength of controls and the relative riskassociated with particular programs and operationsmust be ensured.

Nowadays Many companies, researchers and cloud developers are working on clouds and most of them work spatially on cloud security as the biggest challenge of like Amazon, Google, IBM and so on.They design their methods and publish them. Also they alwaystest their new method on cloud systems or even big social networks but still they don't find a complete way to create a secure cloud. Some organizations like ENISA, CSA and ISAKA survey the future of cloud security.

We think our designed model has more secure levels than other models that can make clouds more secure. But we don't claim our model is complete because several critical pieces of technology, such as a solutionfor federated trust, are not yet fully realized, impingingon successful deployments. In security issues completeness is an ultimate goal but no one can access it.

REFERENCES

[1] Wayne A. Jansen, ―Cloud Hooks: Security and Privacy Issues in Cloud Computing‖, 44th Hawaii International Conference on System Sciesnces 2013.

[2] D. Hubbard and M. Sutton, “Top Threats to Cloud Computing V1.0,” Cloud Security Alliance, 2013. Available:http://www.cloudsecurityalliance.org/topthreats

[3] P. Mell, T. Grance, The NIST Definition of Cloud Computing, Version 15, National Institute of Standards and Technology, October 7, 2011,http://csrc.nist.gov/groups/SNS/cloud-computing

[4] http://www.openstack.org/projects/openstack-security/

[5] L. Youseff, M. Butrico, D. D. Silva, Toward a Unified Ontology of Cloud Computing, Grid Computing Environments Workshop, held with SC08, November 2014 http://www.cs.ucsb.edu/~lyouseff/CCOntology/CloudOntolo gy.pdf

[6] M. Jensen, J. Schwenk, N. Gruschka, and L. Lo Iacono, "On technical security issues in cloud compxuting," in Proceedings of the IEEE International Conference on Cloud Computing (CLOUD-II), 2012.

[7] D. Cappelli, A. Moore, R. Trzeciak, T. J. Shimeall, Common Sense Guide to Prevention and Detection of Insider Threats, 3rd Edition, Version 3.1, CERT, January 2015, http://www.cert.org/archive/pdf/CSG-V3.pdf

[8] M. Burkhart, M. Strasser, D. Many, and X. Dimitropoulos,"SEPIA: Privacy-Preserving Aggregation of Multi-Domain Network Events and Statistics," in USENIX Security Symposium, 2013.

[9] Y. Keleta, J. H. P. Eloff, H. S. Venter, Proposing a Secure XACML Architecture Ensuring Privacy and Trust, Research in Progress Paper, University of Pretoria, 2005, http://icsa.cs.up.ac.za/issa/2005/Proceedings/Research/093_A rticle.pdf

[10] http://docs.openstack.org/trunk/openstack/compute/admin/content/components-of-openstack.html

[11] S. Ramgovind, M.M. Eloff, and E. Smith, "The Management of Security in Cloud Computing," IEEE, 2010,

[12] X. Jing, and Z. Jian-jun, "A brief Survey on the Security model of Cloud Computing," IEEE, 2013

[13] M. P. Eisenhauer, Privacy and Security Law Issues in Off-shore Outsourcing Transactions, Hunton& Williams LLP, The Outsourcing Institute, February 15, 2012, http://www.outsourcing.com/legal_corner/pdf/Outsourcing_Privacy.pdf

[14] B. R. Kandukuri, R. Paturi V, A. Rakshit, Cloud Security Issues, IEEE International Conference on Services Computing, Bangalore, India, September 21-25, 2015

[15] S. Overby, How to Negotiate a Better Cloud Computing Contract, CIO, April 21, 2010, http://www.cio.com/article/591629/How_to_Negotiate_aBetter_Cloud_Computing_Contract

[16] T. Ristenpart, E. Tromer, H. Shacham, S. Savage, Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds, ACM Conference on Computer and Communications Security, November 2014

[17] C. Wang, "Forrester: A Close Look At Cloud Computing Security Issues," CSO. 2009

[18] J. Somorovsky, M. Heiderich, M. Jensen, J. Schwenk, N. Gruschka, and L. Lo Iacono, "Breaking the clouds – security analysis of cloud management interfaces," (in submission),2014.

[19] S. Pearson, Taking Account of Privacy when Designing Cloud Computing Services, ICSE Workshop on Software Engineering Challenges of Cloud Computing, May 23, 2013, Vancouver, Canada

[20] A. Greenberg, IBM's Blindfolded Calculator, Forbes Magazine, July 13, 2014

INTENTIONAL BLANK

DETERMINING THE CORE PART OF SOFTWARE DEVELOPMENT CURRICULUM APPLYING ASSOCIATION RULE MINING ON SOFTWARE JOB ADS IN TURKEY

Ilkay Yelmen[1] and Metin Zontul[2]

[1]Department of Computer Engineering,
Istanbul Technical University, Istanbul, Turkey
yelmen@itu.edu.tr
[2]Department of Software Engineering,
Istanbul Aydin University, Istanbul, Turkey
metinzontul@aydin.edu.tr

ABSTRACT

The software technology is advancing rapidly over the years. In order to adapt to this advancement, the employees on software development should renew themselves consistently. During this rapid change, it is vital to train the proper software developer with respect to the criteria desired by the industry. Therefore, the curriculum of the programs related to software development at the universities should be revised according to software industry requirements. In this study, the core part of Software Development Curriculum is determined by applying association rule mining on Software Job ads in Turkey. The courses in the core part are chosen with respect to IEEE/ACM computer science curriculum. As a future study, it is also important to gather the academic personnel and the software company professionals to determine the compulsory and elective courses so that newly graduated software developers can easily adapt to the software projects in the market without taking extra training.

KEYWORDS

Association Rule Mining, IEEE/ACM Computer Science Curriculum, Software Development Curriculum, Software Job Ads

1. INTRODUCTION

There are many departments that give education in software development in bachelor degree, such as Software Engineering, Computer Engineering, Computer Science or Mathematics Computer. All graduates in the market generally use software engineer, software specialist or software developer titles in Turkey. Actually, software engineering is used instead of software development for the most of time. Therefore, it is very vital to determine the common core part of the curriculums of these departments. Software development is an engineering practice that includes the topics such as design, implementation and maintenance [1]. In the last 30 years, the importance of software development has increased and it has been growing continuously [2].

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 93–106, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60209

This increasing importance and rapid change forced the software education to be adaptive to the market needs. Increasing costs in the software industry, applied wrong strategies, the desires to rise in quality and performance issues and so fast technology changes had revealed the need to educate experts in the field and qualified software developers. Therefore, the university-industry collaboration has gained utmost importance.

Zhengyu stated that a lot of strong software professionals were urgently required in the community but employers felt that the graduates had the software talent shortage while a considerable number of graduates could not find a suitable position [3].

In another study, Kuang and Han proposed the methods of teaching reform as guided by market demand, to update the teaching content, to optimize the teaching methods, to reform the teaching practice, to strengthen the teacher-student exchange and to promote teachers and students together because software development training could not meet the needs of the community [4].
Among the software development departments, software Engineering (SE) is the fastest-evolving engineering discipline that has ability to provide tools and methods for all areas of society [5]. This situation increases the responsibility of SE education to prepare SE professionals for the industry by providing them with skills to meet the expectations of the software industry. Innovations and improvements in the curriculum are required to bridge academia-industry gap [6] since SE education has inability to provide students with large-scale software development experiences [7]. However, only universities can produce highly skilled professionals who can satisfy the needs of software industry by taking into account different standards, frameworks and recommendations developed by interest groups [8].

A study presented software engineering education evolvement in Turkey to provide an assessment of SE curriculum in Turkish Universities with respect to IEEE/ACM guidelines given in SEEK (2004) and to provide a guideline to universities conducting an SE programme at undergraduate level to align their course curriculum with IEEE/ACM guidelines [9].

Students should have necessary background of programming experience for the study of software engineering concepts in their curriculums. In order to satisfy this condition, the current software engineering guidelines include concepts and programming paradigms that must be mastered through study and practice. The well-known guideline for software engineering curricula is recommended by IEEE/ACM. This guideline gives the standards related to course scheduling, faculty preparation, student loads, hardware and software resources, instructional materials and curriculum development. ACM published “Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering (SE2004)” to provide guidance to academic institutions and accreditation agencies about what should constitute an undergraduate software engineering education [10].

This study is related to curriculum development and human resources in software development. In addition, association rule mining on software job ads is applied. There are similar studies in literature as follows. Chien and Chen developed a data mining framework for personnel selection to explore the association rules between personnel characteristics and work behaviors, including work performance and retention. Moreover, they used decision tree analysis to discover latent knowledge and extract the rules to assist in personnel selection decisions [11].

Mohsin, Ahmad, Din, Mahamud and Din proposed an intelligent model that is aimed at facilitating key workers select suitable trainees for a training program. In this study, trainees dataset was mined using association rule to discover important personality characteristics. Their model produced an efficient selection process and suitable trainees [12].

In another study, Ali and Rajamani presented the solution for selecting appropriate talented personnel resumes without risk factors using association rule mining. The practical experimental results obtained from the proposed model encouraged human resource department to take prompt decisions for recruiting talented personnel accurately without wasting interviewers' time of employer and employee. Also, they indicated that the proposed system reduced frequent resignations, improved performance of talented personnel without training cost and continuous monitoring [13].

Finally, Smith and Ali indicated that today's rapid changing and competitive environment required educators to stay alongside of the job market in order to prepare their students for the jobs being demanded. They also implied that data mining methods were suitable for this kind of analysis due to the large volume of job data generated through the web instead of the classical data analysis methods. Their study illustrated the experience with employing mining techniques to understand the trend in IT Technology jobs. At the end, collected data from an online agency was analysed to reach a conclusion about the trends in the job market [14].

In this study, the core part of Software Development Curriculum is determined by applying association rule mining on Software Job ads in Turkey. As a result, software engineering or related fields that give education in software development should include these core courses in their curriculum in order to adapt the software development industry in Turkey.

The rest of this paper is organized as follows: The second chapter makes mention of association rules and the third chapter depicts how to apply association rule mining on software ads. The fourth chapter gives results and discussion with respect to the association rule mining. Finally, the fifth chapter gives the conclusion.

2. ASSOCIATION RULES

One of the important tasks for Knowledge Discovery in data is Association Rule Mining which is a well-known procedure in data mining. In its basic structure, every association rule fulfilling the minimum support and confidence are extracted [15]. The general purpose of an association rule A => B is to denote that records possessing attribute A also tend to possess attribute B. The aim is to find association rules which are considered sufficiently interesting as defined by one or more measures. Most common formulas for support and confidence are as follows [16]:

$$\text{Support}(A \Rightarrow B) = \frac{||A \wedge B||}{|D|} \quad (1)$$

$$\text{Confidence}(A \Rightarrow B) = \frac{||A \wedge B||}{||A||} \quad (2)$$

where $|D|$ indicates total number of records and $||A||$ refers to total number of record including A.

2.1. Apriori Algorithm

Apriori is an algorithm which is developed for common set learning mining on transactional database and association rule learning [17]. Apriori uses a level-wise search, where k-itemsets are used to explore (k+1) itemsets. First, the set of frequent 1-itemsets denoted by L1 is found by scanning the dataset to find the count for each item, and collecting those items satisfying minimum support. Then, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the dataset. The Apriori property is used to improve the efficiency of the level-wise generation of frequent item sets by reducing the search space [18].

The Apriori property is based on the following interpretations. By definition, if an item set I does not satisfy the minimum support threshold, min_sup, then I is not frequent (P(I)<min_sup). If an item A is added to the item set I, then the resulting item set cannot happen more frequently than I. Therefore, I U A is not frequent either (P(I U A) < min_sup) [18].

3. ASSOCIATION RULE MINING ON SOFTWARE JOB ADS

3.1 Collecting Data

The data used in this study is taken from one of the popular job recruitment site in Turkey and currently available ads about software are examined from 5 big cities that are Istanbul (European Side), Istanbul (Asian Side), Ankara, Izmir and Bursa. At the end feature list and sub-categories are prepared with respect to job ads for using in the data set [19].

653 software job ads in 5 big cities are collected from this job recruitment site in Excel format. Finally, 30 main features are determined for finding suitable candidates for the position as shown in Tab. A.I in Appendix.

3.2 Data Processing

After creating the initial dataset, it is observed that the year of experience in software features can be important. Consequently, 30 features and some information within these features have been separated. For example, considering the experience in programming languages on the job recruitment site, programming languages are divided into 8 pieces as C, C++, C#, Java, Php, Objective C and the other programming languages. On the other hand, experience levels are divided into 7 as 0 (Not Acknowledged), 1, 2, 3, 4, 5 and 5+ (Years). Programming language and experience level features are combined for simplicity. For example, if C# feature contains 2, at least 2 year-experience is required. In addition, as example qualifications in the special programming techniques are grouped and new features are created under new names as Software Architecture Methodologies, Software Patterns, Programming Paradigms and Other Software Development Processes as shown in Tab. A.II in Appendix. Finally, 653 ads are entered for the features in Table A.II.

Then, the data is visualized in terms of database systems, programming languages, Front-end technologies and other software technologies. As shown in Fig. 1, the leading database systems in

job ads are MS SQL, Oracle and MySql. The mostly used programming languages are C#, Java and C++ in job ads as depicted in Fig. 2

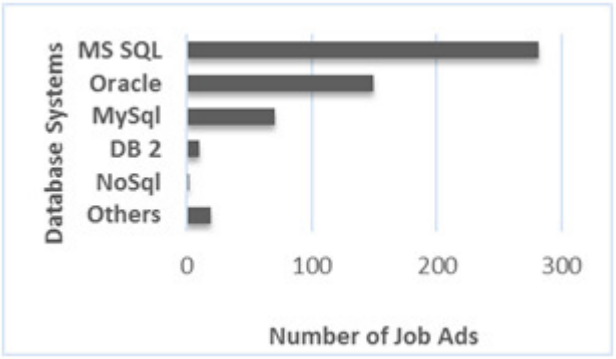


Figure 1. The leading database systems in software job ads

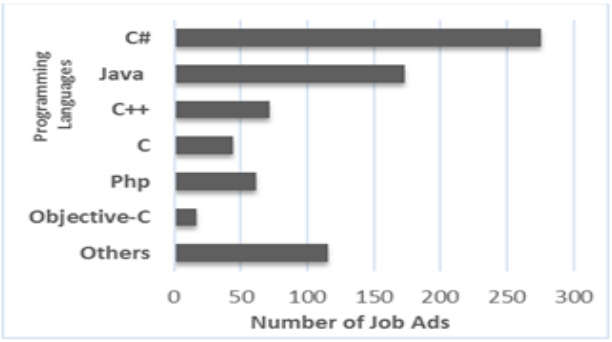


Figure 2. The leading programming languages in job ads

In similar way, Fig. 3 shows the leading front-end development technologies in job ads where Javascript is prominent technology for front-end development. The other software technologies such as software architecture, software paradigms and web services are essential as shown in Fig. 4.

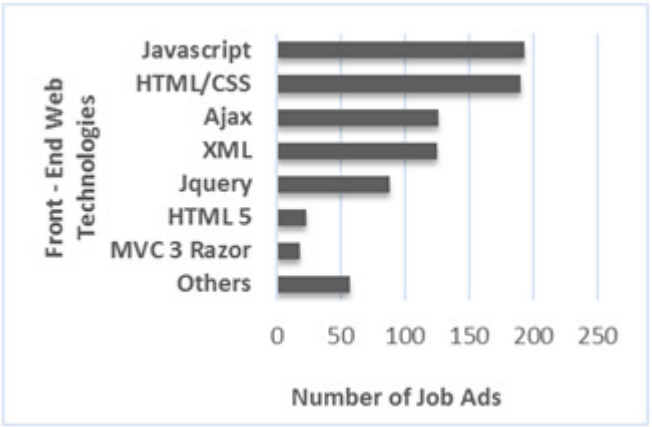


Figure 3. The leading Front-end technologies in job ads

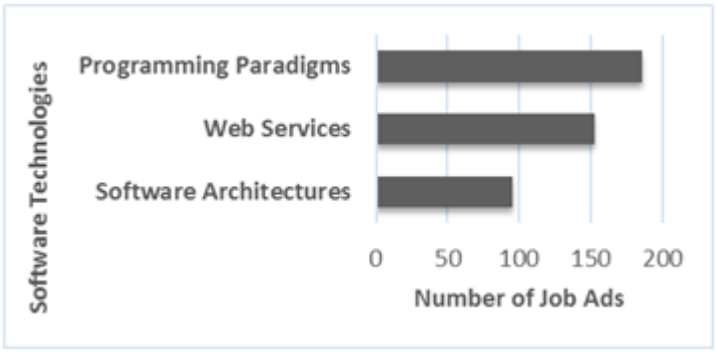


Figure 4. The leading software technologies in job ads

3.3 Applying Apriori Algorithm

Fig. 1, Fig. 2, Fig. 3 and Fig. 4 above give the frequencies of related technologies but they don't give which technologies are related to each other. Moreover, one job ad can contain more than one software technology at the same time. For this purpose, association rule mining by using apriori algorithm is applied in assessment. Different combinations trying out through the features analysis has been done. Minimum support value has been taken as 0.1 and confidence value has been taken as 0.5 in the analysis process. At the end, 54 rules are determined regarding software development.

4. RESULTS AND DISCUSSION

The rules obtained by apriori algorithm are divided into 3 parts as follows:

1. Programming languages, frameworks and databases

2. Front-end web technologies

3. Web services, software architectures and programming paradigms

There are 23 rules related to programming languages, frameworks and databases as shown Tab. 1. If the result sides of these rules are noted, Education_Level=3 (B.Sc. Degree), Position=1 (Software Specialist) are prominent results. On the left side of rule 1, {Java=0} condition means that Java is necessary but the year of experience is not important. The same thing can be said for C# if the rule 8 is considered. While 87% of ads that want Java experience require at least B.Sc. degree (Rule 1), 74% of ads that want C# experience require at least B.Sc. degree (Rule 14). It can be argued from these rules that the university degree is more important in Java than C#. Looking at rules 6 and 12, the similar result can be obtained for Oracle (79%) and MS SQL Server (75%). The rules 4 (80%), 5 (80%) show that the ads seeking for C# experience with MS.NET or ASP.NET framework experience categorize the job seekers as Software Specialists. In rule 2, the same result can be obtained for 82% of ads seeking for C# experience with MS SQL Server experience. From the rule 4, 5 and 7, it can be concluded that Job seekers knowing C# should have enough experience in MS.NET, ASP.NET framework and MS SQL Server. If the rule 8 is compared with the rule 21, it can be said that while 77% of the ads seeking for C#

categorize the job seekers as Software Specialist, the corresponding ratio for Java is 67%. From these rules, it can be decided that C# experience more valuable for the software firms in Turkey. The similar result is valid for MS SQL Server (75% in rule 13) and Oracle (71% in rule 18). Another interesting result can be obtained from the rules 22 and 23 that Oracle ads and Java ads are separated from each other. However, C# and MS SQL Server are combined in rules 2 and 15. It means that Oracle and Java are complicated technologies that cannot be known by one specialist while C# and MS SQL Server are moderate technologies that can be known by one specialist.

Table 1. Rule Extraction for Programming Languages Frameworks and Database through the Apriori Algorithm

No	Rule	Confidence Value
1	If {Java=0} ==> Education_Level =3	0.87
2	If {C#=0 ∩ MSSQL=0} ==> Position =1	0.82
3	If {Other_Languages =0} ==> Position=1	0.81
4	If {C#=0 ∩ AspDotNetFramework=0} ==> Position =1	0.80
5	If {C#=0 ∩ DotNetFramework=0} ==> Position =1	0.80
6	If {Oracle=0} ==> Education_Level =3	0.79
7	If {AspDotNetFramework =0} ==> C#=0	0.78
8	If {C#=0} ==> Position=1	0.77
9	If {DotNetFramework=0} ==> Position =1	0.76
10	If {DotNetFramework=0} ==> Education_Level =3	0.75
11	If {AspDotNetFramework=0} ==> Position =1	0.75
12	If {MSSQL=0} ==> Education_Level =3	0.75
13	If {MSSQL=0} ==> Position =1	0.75
14	If {C#=0} ==> Education_Level=3	0.74
15	If {C#=0 ∩ MSSQL=0} ==> Education_Level =3	0.73
16	If {C#=0 ∩ DotNetFramework=0} ==> Education_Level =3	0.73
17	If {Other_Languages=0} ==> Education_Level =3	0.72
18	If {Oracle=0} ==> Position =1	0.71
19	If {DotNetFramework =0} ==> C#=0	0.68
20	If {AspDotNetFramework=0} ==> Education_Level =3	0.67
21	If {Java=0} ==> Position =1	0.67
22	If {Java=0} ==> Education_Level =3 ∩ Position =1	0.60
23	If {Oracle=0} ==> Education_Level =3 ∩ Position =1	0.56

20 rules are available about front-end web technologies as shown in Tab. 2. The rule 24 indicates that 83% of ads that want ajax knowledge require javascript as well (year of experience is not important). Moreover, 81% of ads that include ajax and software specialist together want javascript, too. It can be understood from the rules 24 and 25 that ajax and javascript should be taught together. Similar results can be obtained for the rules 26, 27, 28, 30 in a way that there is a strong relationship among ajax, jquery, javascript, HTML and CSS in terms of software development education in front-end web technology. Also, the rules 32, 34, 36, 37 and 39 imply that at least B.Sc. degree is required at most of the times for front-end development. The rules 24 and 43 show that while 83% of ads that want ajax knowledge require javascript, only 53% of ads that want javascript require ajax. From these rules, it can be extracted that javascript education is more fundamental than ajax for front-end education. Another important item for front-end development is XML as indicated in the rules 32, 33, 40 and 41. The rules 32 and 33 indicate that

the candidates knowing XML should have at least B.Sc. degree and be software specialist at an important level over 70%. The rules 40 and 41 imply the relation between XML and other web technologies HTML, CSS and web services. As 62% of ads requiring XML also want HTML and CSS (rule 40), 56% ads requiring XML want web services. These rules mean that XML is more common data format for data transfer in web platforms. Actually, a few ads include JSON data format but they are eliminated by apriori algorithm because of their low support count. It means that XML is more common in the market in Turkey.

Table 2. Rule Extraction for Front-End Web Technologies through the Apriori Algorithm

No	Rule	Confidence Value
24	If {Ajax=0} ==> Javascript=0	0.83
25	If {Ajax=0 ∩ Position =1} ==> Javascript=0	0.81
26	If {Javascript=0} ==> HTML_CSS=0	0.81
27	If {Ajax=0} ==> HTML_CSS=0	0.78
28	If {Jquery=0} ==> HTML_CSS=0	0.78
29	If {Ajax=0} ==> Position =1	0.78
30	If {Jquery=0} ==> Javascript=0	0.76
31	If {Ajax=0 ∩ Javascript=0} ==> Position =1	0.76
32	If {XML=0} ==> Education_Level =3	0.75
33	If {XML=0} ==> Position =1	0.71
34	If {HTML_CSS=0} ==> Education_Level =3	0.70
35	If {Javascript=0} ==> Position =1	0.68
36	If {Ajax=0} ==> Education_Level =3	0.68
37	If {Javascript=0} ==> Education_Level =3	0.67
38	If {Education_Level =3 ∩ Javascript=0} ==> Position =1	0.67
39	If {HTML_CSS=0 Javascript=0} ==> Education_Level =3	0.66
40	If {XML=0} ==> HTML_CSS=0	0.62
41	If {XML=0} ==> Web_Services=0	0.56
42	If {HTML_CSS=0 ∩ Javascript=0} ==> Ajax=0	0.55
43	If {Javascript=0} ==> Ajax=0	0.53

11 rules are obtained related to web services, software architectures and programming paradigms as depicted in Tab. 3. These rules mostly focus on B.Sc. degree and software specialist position. In fact, there are other ads focusing on other positions such as database administrator or software test specialist but they are eliminated because of their low support counts. This means that the most of software firms give ads focusing on software specialists as shown in rules 49, 51, 52. From the other perspective, software architectures, web services and programming paradigms (object oriented programming, functional programming etc.) topics are very special software technologies that should be carried out by software engineers or equivalents having at least B.Sc. degree as seen in rules 44, 45, 46, 47 and 48. The rules 53 and 54 imply B.Sc. degree and software specialist position at the same time. 53% ads including web services imply B.Sc. degree and software specialist position together (rule 53). The same thing is valid for programming paradigms (rule 54). The Programming paradigm stands for the styles of various programming languages such as Phyton, Lisp, F# and Objective-C.

Table 3. Rule Extraction for Web Services, Software Architectures and Programming Paradigms through the Apriori Algorithm

No	Rule	Confidence Value
44	If {Software_Architectures=0} ==> Education_Level =3	0.84
45	If {Web_Services=0} ==> Education_Level =3	0.83
46	If {Position=1 ∩ Web_Services=0} ==> Education_Level =3	0.80
47	If {Programming_paradigms=0} ==> Education_Level =3	0.78
48	If {Position=1 ∩ Programming_paradigms=0} ==> Education_Level =3	0.77
49	If {Software_Architectures=0} ==> Position =1	0.68
50	If {Programming_paradigms=0} ==> Position =1	0.68
51	If {Education_Level =3 ∩ Programming_paradigms =0} ==> Position =1	0.67
52	If {Education_Level =3 ∩ Web_Services=0} ==> Position =1	0.64
53	If {Web_Sevices=0} ==> Education_Level =3 ∩ Position =1	0.53
54	If {Programming_paradigms =0} ==> Education_Level =3 ∩ Position =1	0.53

With respect to the rules above, the following courses should be included in Software Development Curriculum as compulsory core courses as shown in Tab. 4. These courses are compatible with IEEE/ACM computer science curriculum where it is dictated that successfully deploying an updated computer science curriculum at any individual institution requires sensitivity to local needs [20]. The rules above correspond to the local needs in Turkey.

The rules 2, 10, 12 and 15 in Tab. 1 imply that MS SQL Server is a fundamental database system for software developers. Thus, Database Systems-I course focusing on relational database concept by MS SQL Server is included in the curriculum. In similar way, since Oracle is also very popular in the market according to the rules 6, 18, 23 in Tab. 1 Database Systems-II course focusing on Oracle should be involved. The most popular language in Turkey is C# on MS.Net framework according to the rules 4, 5, 7, 8, 9, 10, 15, 16 and 19. Because the easiest way to start to learn C# is Windows Desktop environment, Desktop Programming course should be given in the curriculum. The rules 1, 21, 22, 47, 48, 50, 51, and 54 in Tab. 1 indicate that Java language is second popular language in Turkey. Since Java includes all object-oriented principles Object-Oriented Programming course applying the basic principles on Java is recommended. Web development can be divided into two parts as front-end and back-end web development. While the rules 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42 and 43 in Tab. 2 imply Front-End Web Development course, the rules 4, 7, 11, 16, 19 in Tab. 2 involve Back-End Web Development. Front-end development should include HTML, CSS, Javascript, Jquery, Ajax, XML. Relating to the rules 4, 7, 11, 16, 19 in Tab. 1, the most popular back-end web programming is ASP.NET with C# in Turkey. As for the rules 47, 48, 50, 51, 54, the programming paradigm concept refers to various programming languages having different programming styles. As a result, Programming Language Concept lecture should be included in order to teach different programming languages such as Phyton, Lisp, F#, Objective-C. Finally, Software Architecture course is recommended as compulsory course since the rules 44 and 49 in Tab. 3 show the importance of software architectures focusing on multi-tier architectures and web services.

Table 4. The Core Part of Software Development Curriculum for Undergraduate Degree Programs

Year/ Semester	Related Tech.	Course Name	Related Rules
2/1	MS SQL Server	Database Systems-I	7, 12, 13, 15
2/1	C#, MS.Net Framework	Desktop Programming	5, 7, 8, 9, 10, 14, 15, 19
2/1	Java	Object Oriented Programming	1, 21, 22, 47, 48, 50, 51, 54
2/2	Oracle	Database Systems-II	6, 18, 23
2/2	HTML, CSS, JavaScript, Jquery, Ajax, XML	Front-End Web Development	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43
3/1	ASP.NET	Back-End Web Development	4, 7, 11, 16, 19
3/1	Phyton, Lisp, F#, Objective-C	Programming Language Concepts	47, 48, 50, 51, 54
3/2	Web Services, Multi-Tier	Software Architecture	44, 49

5. CONCLUSION

Due to consistent growth in software market and rapid change in software technology, the adaption of software development curriculum is necessary with respect to criteria desired by software development industry.

This study has a contribution to the literature in a way that it applies association rule mining on software job ads to help the managers decide on the software development curriculum. In this study, the software job ads regarding the first 5 cities having intensive ads are obtained from a famous Turkish employment web site. While determining the features required for association rule mining, the years of experience on software expertise areas is considered. After applying association rule mining, the rules related to desires of software companies are achieved.

Considering the criteria owned by software job ads, it is investigated that practical part of software development education should be increased. Moreover, since the most of job ads seek the employee at minimum B.Sc. level, the importance of software engineering or related fields is increasing as well. As the companies request the new software technologies as well as fundamental programming abilities, it is vital to revise the software development curriculum at the universities. However, the most of them are very late to adapt their curriculums to the criteria of the companies. Actually, it not enough to revise the curriculums but also academic personnel in these departments should renew their knowledge on the new software technologies.

Since there are many departments related to software development such as software engineering, computer engineering etc., it is necessary to determine the core courses of software development. For this purpose, the core courses are determined by using association rule mining on software job ads and IEEE/ACM computer science curriculum. The core courses should be common for all departments related to software development because they reflect the local need of the software development companies in Turkey.

As a future study, it is also important to gather the academic personnel and the software company professionals to focus on the hot software technologies. With respect to the results obtained from these meetings, the compulsory and elective lectures should be determined so that newly graduated software developers can easily adapt to the software projects in the market. As a result, it will be easier to find a job for them and to hire a proper developer for the companies without giving extra training.

REFERENCES

[1] Laplante, P. A.. What every engineer should know about. Taylor & Francis Group., Boca Raton, FL, 2007 [Online]. Available:http://lib.mdp.ac.id/ebook/Karya%20Umum/Every-Engineer-Should-Know-about-Software-Engineering.pdf

[2] Mccracken, M., et al. A proposed curriculum for an undergraduate software engineering degree. In: Software Engineering Education Training. // In Proc. 13th Conference on. IEEE, 2000, p. 246-257.

[3] Zhengyu, G. R. Y. Strengthening Practices and Researches in the Education of Corporation of Enterprises, Colleges and Institutions. Jiangsu Social Sciences, S2, 2007.
Available: http://en.cnki.com.cn/Article_en/CJFDTOTAL-JHKX2007S2007.htm

[4] Kuang, L. Q.; Han, X. The Research of Software Engineering Curriculum Reform. Physics Procedia. [Online] 33(2012), pp. 1762-1767.
Available: http://www.sciencedirect.com/science/article/pii/S1875389212015957

[5] Král, J.; Zemlicka, M. Engineering Education-A Great Challenge to Software Engineering. // In Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference, pp. 488-495.

[6] Shaw, M.; Herbsleb, J. D.; Ozkaya, I. Deciding What to Design: Closing a Gap in Software Engineering Education. // Invited paper for Education and Training Track of 27th Int. Conf. on Software Engineering (ICSE 2005), 2005, pp. 607 – 608.

[7] Su, H.; Jodis, S.; Zhang, H. Providing an integrated software development environment for undergraduate software engineering courses. // Journal of Computing Sciences in Colleges. [Online] 23, 2(2007), pp. 143-149. Available: http://dl.acm.org/citation.cfm?id=1292453

[8] Jaakkola, H.; Henno, J.; Rudas, I. J. IT Curriculum as a complex emerging process. // In Computational Cybernetics, 2006. ICCC 2006. IEEE International Conference on 2006, pp. 1-5

[9] Mishra, A.; Yazici, A. An Assessment of the Software Engineering Curriculum in Turkish Universities: IEEE/ACM Guidelines Perspective. // Hrvatski časopis za odgoj i obrazovanje. [Online] 13, 1(2011), pp. 188-219. Available: http://hrcak.srce.hr/72403

[10] Computing Curricula (CC) (2005). Guidelines for Associate-Degree Transfer Curriculum in Software Engineering.
Available:http://www.capspace.org/committee/CommitteeFileUploads/TYC_SE_report.pdf

[11] Chien, C. F.; Chen, L. F. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. // Expert Systems with applications. [Online] 34, 1(2008), pp. 280-290. Available: http://www.sciencedirect.com/science/article/pii/S0957417406002776

[12] Mohsin, M. F.; Ahmad, M. F; Din, A. M.; Mahamud, K.R.K.; Din, R. An intelligent trainee selection model. // In Computer Sciences and Convergence Information Technology (ICCIT), 6th International Conference on, 2011, pp. 390-393.

[13] Ali, M. M.; Rajamani, L. Automation of decision making process for selection of talented manpower considering risk factor: A data mining approach. // In Information Retrieval & Knowledge Management (CAMP), International Conference on 2012, pp. 39-44.

[14] Smith, D.; Ali, A. Analyzing computer programming job trend using web data mining. // Issues in Informing Science and Information Technology. [Online] 11, 2014, pp. 203-214. Available: http://iisit.org/Vol11/IISITv11p203-214Smith0494.pdf

[15] Ishibuchi, H.; Kuwajima, I.; Nojima, Y. Prescreening of candidate rules using association rule mining and Pareto-optimality in genetic rule selection. // In Knowledge-Based Intelligent Information and Engineering Systems. [Online] 4693, 2007, pp. 509-516.
Available: http://link.springer.com/chapter/10.1007/978-3-540-74827-4_64

[16] Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. // In ACM SIGMOD Record. 22, 2(1993), pp. 207-216.

[17] Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. // In Proc. 20th int. conf. very large data bases. VLDB 1215, 1994, pp. 487-499.

[18] Han, J.; Kamber, M. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann. 2006.

[19] Job Recruitment Website (JRW). 2014. Available: http://www.kariyer.net/

[20] Curriculum Guidelines for Undergraduate Degree Programs in Computer Science (CGCS) (2013, December). Computer Science Curricula 2013 [Online].
Available: http://www.acm.org/education/CS2013-final-report.pdf

AUTHORS

Ilkay Yelmen received the BS degree in Software Engineering from the Istanbul Aydin University in 2013 and he is currently pursuing the MS degree in Computer Engineering at the Istanbul Technical University. His research interests are Data Mining and Natural Language Processing.

Metin Zontul received the BS Degree in Computer Engineering from Middle East Technical University, Ankara, Turkey, the MS Degree in Computer Science from Erciyes University, Kayseri, Turkey, and the PhD Degree in Numerical Methods from Cumhuriyet University, Sivas, Turkey. His research interests are software development, information systems, soft computing and data mining. He is currently a faculty member of Software Engineering Department at Istanbul Aydin University, Istanbul, Turkey.

Appendix

Table AI. Main Features and Their Sub Details in Software Job Ads

Education Level	Delphi	Sybase	Action Script	WAF	DDD	Windows Phone	Hudson	Junit	DirectX
Not acknowledged	Objective-C	Oracle	XSLT	WWF	Comet	Blackberry	PMP	iBATIS	JSON
High School	Progress-ABL	MS Sql Server	Vb Script	Data Mining	MVC	**Military Status**	OCP	Doctrine ORM	Dojo
Associate (Student)	TROIA	Neo4j	DHTML	OLAP	SDLC	Not acknowledged	CEH	Spring MVC	Google Closure
Associate (Graduate)	Ruby	PostgreSQL	XLS	Primefaces	**Web Services**	Postponed	**Framework**	Django	Prototype
B.Sc. (Student)	**Foreign Lang. Knowledge**	Sql	Xpath	JMX	SOA	Free	PHP5	Magento	Django
B.Sc. (Graduate)	Not acknowledged	**Graduated Area**	XSL-FO	JPA	WS	**Certificate Info**	Struts	**Travel Ban**	OpenGL
M.Sc. (Student)	English	Mathematics	DOM	EJB	SOAP	Not acknowledged	Servlets	Not acknowledged	RUP
M.Sc. (Graduate)	German	Software Eng.	Xquery	JDO	WSDL	MCSD	Spring	Asked	MFC
Phd (Student)	**Work Experience**	Computer Eng.	XSD	JPA	UDDI	MCT	Hibernate	**Smoking Status**	ODM
PhD (Graduate)	Not acknowledged	Electrical and Electronics Eng	XSL	JTA	RESTful	MCPD	Coherence	Non-use	Enterprise
Operating Systems	1	Industrial Eng.	XAML	JCA	**Position**	ISTQB	MVC	Unspecified	**Social API**
Not acknowledged	2	Electronics and Comm. Eng.	**Technical Skills**	JAXB	Software Architect	CSTE	.NET	**Age Information**	Facebook
Linux	3	Management Inf. Systems	Not acknowledged	JAX-RPC	Soft. Test Specialist	CISM	Entity	Not acknowledged	Twitter
Unix	4	Statistics	Looking for	JSF	Junior Software Dev	MCDBA	Zend	20-25	OpenGraph
Windows OS	5	Physics	**Personal Skills**	CRM	Senior Software Dev	PMI	Codeigniter	25-30	**Auxiliary Platforms**
Macintosh	6	Mathematics-Computer	Not acknowledged	ERP	Medior Software Developer	MCAD	Symfony	30+	Glassfish
Office Applications	7	Computer Programming	Looking for	**Special Programming Technics**	Mobile Soft. Develop. Specialist	CSTE	CakePHP	**IDE**	SAP Basis
Not acknowledged	8	Mathematics Eng.	**Position Type**	N-tier	Project Manager	ISEB	Agavi	Eclipse	IIS
Asked to know	9	Comp. and Ins.Tech. Edu.	Part Time	UML	Business Analyst	MCP	Cocoa Touch	Netbeans	Git
Programming Languages	10	Electrical Eng.	Full Time	MVVM	Analyst Developer	CCSE	UIKit	Visual Studio	GitHub
C#	10+	Electronics Eng.	Intern	MVC 3 Razor	Game Programmer	TMAP	Sencha	Dev C++	SVN
Java	**Graphic Applications**	Engineering	Volunteer	OOP/D/A	Team Leader	**Test Automation and Management Tools**	ExtJS	MATLAB	SSIS
C	Not acknowledged	Not acknowledged	**Special Software Tech.**	Multithreading	Automation Project Engineer	Jira	Android	Xcode	SSAS
C++	Asked to know	**Web Technologies(Client-Side)**	ORM	MVP	Database Specialist	Selenium IDE	OAF	Emacs	SSRS
PHP	**Database**	HTML	WCF	Middleware	Software Developer	Apache Jmeter	EJB	Vi	Mercurial
Phyton	T-Sql	CSS	WPF	Nhibernate	Web Developer	SOAP UI	Silverlight	Borland	TFS
ABAP	Pl-Sql	Ajax	Asp.Net	SEO	**Gender**	Mantis	RichFaces	Zend Studio	Mercurial
Scala	Progress	Jquery	ASP	TDD	Not acknowledged	Bugzilla	ADF	Adobe Air	TFS
Axapta(X++)	DB2	Javascript	RSS	SCRUM	Male	ClearQuest	Play!	CVS	Developpress
COBOL	My Sql	XML	MS Azure	LINQ to SQL	Female	Mercury	Cross-platform	**Libraries**	LAMP
Visual Basic	Mongodb	XHTML	JMS	ALM	**Mobile Platforms**	Firebug	Apache cxf	ExtJs	JMS
Perl	No-SQL	HTML5	J2EE	MDD	IOS	Firebug Lite	Toplink	Tkinter	Nginx
Unity	Sql Lite	CSS3	Apache Tomcat	Cryptology	Android	JIRA	PhoneGap	QT	Swing

Table AII. The Final Dataset Structure for Software Job Ads

Education Level			**Position Type**	20-25			
Not acknowledged			Part Time	25-30			
High School		Not acknowledged	Full Time	30+			Not acknowledged
Voca. High School		1	Intern	**Certificate Info**			1
B.Sc.	**C#**		Volunteer	Not acknowledged			
M.Sc.	**C**		**Military Status**	Analysis			
PhD	**PHP**	2	Not acknowledged	Project Manage.	**Android**	**UIKit**	2
Experience	**Java**		Postponed	Development	**IOS**	**Cocoa Touch**	
Not acknowledged	**C++**		Free	Database	**Blackberry**	**IOS/OS Other Frameworks**	
1	**Objective - C**	3	**Position**		**Windows Phone**	**Other Frameworks**	3
2	**Other Languages**			Security	**Personal Skills**	**Javascript Library**	
3	**Oracle**		Not acknowledged	Test	**Technical Skills**	**GUI Library**	
	MS SQL		Software Architect (Software Specialist)	**Gender**	**Visual Studio**	**Other Libraries**	
	DB2				**Eclipse**	**Facebook**	
4	**NO-SQL**	4		Not acknowledged	**Xcode**	**Twitter**	
5	**MySQL**		Software Test Specialist		**Netbeans**	**OpenGraph**	
6	**Sql Lite**		Junior Software Dev.	Important	**Zend Studio**	**Json**	4
7	**Other Databases**		Senior Software Dev.		**Dev C++**	**Software Architectures**	
	HTML/CSS	5	Medior Software Developer		**Other IDE**	**Methodologies**	
	Ajax		Mobile Software Dev. Specialist		**.NET**	**Software Patterns**	
8	**Jquery**				**ASP.NET MVC Framework**	**Prog. Paradigms**	
9	**HTML5**		Project Manager		**Entity Framework**	**Other Soft. Dev. Processes**	
10	**Javascript**	5+	Business Analyst		**ASP.NET Framework**	**API**	
10+	**MVC 3 Razor**		Analyst Developer		**.NET Other**	**Software tech.**	5
Foreign Lang. Knowledge	**XML**		Game Programmer		**Struts**	**Process**	
Not acknowledged	**Other Web Tech.**		Web Developer		**Servlets**	**Server**	
Elementary	**The Field graduated**	Statistics	Automation Project Eng.		**Spring**	**Source Control**	
Intermediate	Not acknowledged	Physics	Database Specialist		**Hibernate**	**Other Special Prog. Tech.**	
Advanced	Software Engineering	Mathematics-Computer	**Travel Ban**		**Junit**	**Services**	5+
Operating Systems	Computer Eng.	Computer Programming	Not acknowledged		**JSF**	**Testing Tools**	
Not acknowledged	Electrical and Electronics Eng.	Mathematics Eng.	Asked		**RichFaces**	**Component**	
Windows OS	Industrial Eng.	Comp. and Ins. Tech. Education	**Smoking Status**		**Other Java Fra.**	**Other Auxiliary Platforms**	
Windows - Linux-Unix	Electronics Comm.	Graduates from relevant dept.	Non-use		**PHPS Framework**		
		Electrical Eng.	Unspecified		**Zend Framework**		
			Age Info		**Other PHP Fra.**		
Windows - Linux - Unix - MAC OS	Management Information Systems	Electronics Eng.			**Android Framework**		
		Mathematics	Not acknowledged		**Other Android Fr.**		

COMPARATIVE EVALUATION OF FOUR MULTI-LABEL CLASSIFICATION ALGORITHMS IN CLASSIFYING LEARNING OBJECTS

Asma Aldrees[1] and Azeddine Chikh[2] and Jawad Berri[3]

Information System Department,
College of Computer and Information Sciences
King Saud University, Riyadh, Kingdom of Saudi Arabia
[1]asma.aldrees@gmail.com
[2]az_chikh@ksu.edu.sa
[3]jberri@ksu.edu.sa

ABSTRACT

The classification of learning objects (LOs) enables users to search for, access, and reuse them as needed. It makes e-learning as effective and efficient as possible. In this article the multi-label learning approach is represented for classifying and ranking multi-labelled LOs, whereas each LO might be associated with multiple labels as opposed to a single-label approach. A comprehensive overview of the common fundamental multi-label classification algorithms and metrics will be discussed. In this article, a new multi-labelled LOs dataset will be created and extracted from ARIADNE Learning Object Repository. We experimentally train four effective multi-label classifiers on the created LOs dataset and then, assess their performance based on the results of 16 evaluation metrics. The result of this article will answer the question of: what is the best multi-label classification algorithm for classifying multi-labelled LOs?

KEYWORDS

Learning object, data mining, machine learning, multi-label classification, label ranking.

1. INTRODUCTION

The advancement and increasing availability in Internet technologies have changed many activities in life. One of the important activities is Learning which is being supported by these various technologies. The form of online distance learning is gaining a strong attention by learners of all ages with different interests. Learners have found digital learning media to be extremely convenient while learning as it involve the various human senses and different cognitive activities. It is the combination of the web and learning.

E-learning has emerged as a promising domain to facilitate and enhance learning through information technologies. Gerard (2006) [1] suggested that course units in computer-based instruction could be made smaller and combined in various ways for customization and use by

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 107–124, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60210

learners. Learning objects (LOs) are an application of this type of course-units, and through the past years, they have gained the attention in the education area. Nowadays, LO is a concept used very often in different domains regarding learning management systems where it can be described as an essential, major unit that can be shared, reused and retrieved.

LOs should be tagged with metadata description and stored in a digital library, called Learning Object Repository (LOR), for future reuse. . Within the huge number of LOs, the demand to identify and classify them has arisen and become a critical issue in e-learning in order to make it faster and easier to the learners. To achieve this classification, each LO must be tagged with metadata about it to be easily located and later retrieved from repositories. These metadata are descriptive information of the LO, such as its topic, type, and keywords, that allow easy search of LOs.

LOs are mainly annotated with multiple labels, so we would like to be able to retrieve LOs based on any of the associated tags, not only one tag. Therefore, the single-label classification cannot model this multiplicity.

The focus of this paper is on multi-label classification methods [2] [3] [4] for searching LOs based on their tagged metadata. It aims to offer a sorting system that allows recovering and classifying LOs and offering individualized help based on choosing the best and effective classification technique for them.

A second contribution of this paper is creating a new multi-label dataset within a vast number of LOs and their associated metadata from one of the available repositories. The labels in this dataset are automatically generated as metadata and assigned to the LOs,. Metadata generation is a research field, which has been heavily worked on, in the recent years. This contribution will be explained in details in the next sections.

This paper is structured as follows: section 2 explains the main concepts and characteristics that establish LOs as the critical base within the context of web-based learning. Section 3 presents the background material on the multi-label learning, including the: classification techniques and evaluation measures. Also, in this section we will select the effective techniques to be used and compared in this experiment. Section 4 provides the details of the dataset used in this paper In Section 5; we will show the experimental results of comparing the adopted four multi-label classification techniques. Finally, conclusions and future work are drawn in Section 6.

2. CONTEXT OF LEARNING OBJECTS

The concept of LO has received considerable attention, for the first time, and described in 1967 by Gerard [5]. The term LO derived from the idea of Object Oriented Programming (OOP), in which, the parts of code are reused for multiple software applications. This concept suggests that, the ideal way to build a program is to assemble it from standardized, small, interchangeable chunks of code [6].

E-learning is defined as "learning facilitated and supported through the use of Information Technology (IT)". An E-learning experience is made up of the interaction of a number of learning components such as: courses, assessments, teaching materials, study materials, etc.
LOs are a relatively new way of presenting these learning contents. The idea appears to have a

transformation from traditional, direct instruction courseware design approaches, to a more effective and economical strategies for management and reuse of learning resources in computer-based networked environments.

The functionality of LOs can be described as [7]: "Firstly breaking educational material down into modular 'chunks' (objects), where each object can then have its defining properties described (or tagged) using metadata constructs".

Examples of LOs include: multimedia content, instructional content, learning objectives, instructional software and software tools, and persons, organizations, or events referenced during technology supported learning [8].

Recently, many research efforts concentrated on defining LOs. Currently, it appears difficult to arrive at a single definition of a LO that would align communities with diverse perspectives.

To aggregate up what a LO is, we summarize the general specifications of LO among all definitions:

- LOs are a new way of thinking about learning content. Conventionally, content comes in a several small chunks. LOs are smaller units of learning, which indicates that LO is a small component of the lesson.

- LOs are self-contained - each LO is independent, which means that each LO can be considered particularly without connection to other LO.

- LOs are reusable - a single LO may be used in multiple contexts for multiple purposes. That means the LO is the basis for a new LO or expands existing ones.

- LOs can be aggregated - they can be grouped into larger collections of content, including traditional course structures.

- LOs are tagged with metadata - every LO should has descriptive information making it to be easily retrieved. Quite important feature allowing using and reusing LOs.

LOs are annotated and tagged with many metadata descriptions. The most notable standards of metadata for LOs are: the Electrical and Electronic Engineers metadata (IEEE-LOM) [9]; Dublin Core Metadata (DCM) [10]; Instructional Management System (IMS) Global Learning Consortium [11]; Advanced Distributed Learning (ADL) [12]; and Canadian Core Initiative metadata (Can-Core) [13]. Since 2002, LOM has been the standard for describing the syntax and semantics of LOs. It's usually encoded in XML.

The purpose of LOM is to support the reusability, discoverability of LOs and to enable their interoperability. They include the element names, definitions, data types, vocabularies, and taxonomies. LOM focus on the minimum set of features needed to allow the LOs to be searched and managed.

LOs are placed and stored inside LORs, in an attempt to facilitate their reusability so that they can be more easily stored and retrieved on the basis of a description of their content. LORs

support simple and advanced search through the LOs. In simple search, they return the results according to the input keywords given by the user. The advanced search allows the user to specify some specific metadata features to filter LOs in order to meet his specific needs. There are many existing, available LORs, for example, but not limited to; Multimedia Educational Resources for Learning and Online Teaching (MERLOT) [14]; European digital library (ARIADNE) [15]; National Science, mathematics, engineering, and technology education Digital Library (NSDL) [16]; Health Education Assets Library (HEAL) [17]; Education Network Australia (EDNA) [18]; … etc.

In this paper a large dataset will be created, from the ARIADNE repository. It will be composed of a sufficient number of LOs and their related LOM metadata.

3. MULTI-LABLE LEARNING

In the machine learning domain, the traditional single-label classification methods has a large amount of research. These methods are concerned with learning a set of examples that are associated with a single label l from a known finite set of disjoint labels L. However, there is a significant and real problem within the classification, while an example belongs to more than one label. This problem is known as multi-label classification problem. [2,19]. In the multi-label classification, the examples are associated with a set of labels $Y \subseteq L$.

The multi-label learning has two major tasks: multi-label classification (MLC) and multi-label ranking (MLR). In the case of MLC, the idea is to build a predictive algorithm that will provide a list of relevant labels for a given unseen example. On the other hand, the idea in the task of MLR is to provide a ranking of the selected relevant labels for the given unseen example.

Initially, MLC was mainly motivated by application in the domains of text categorization and medical diagnosis. However, nowadays, MLC has attached and is increasingly required by many new application domains, such as semantic annotation of images [20] and video [21]; protein function classification [22]; music categorization into emotions [23]; and Yeast gene functional classification [24].

There are different techniques that have been proposed to be applied to MLC problems, [25]. The next two subsections will describe the common and representative techniques of MLC and their evaluation metrics.

3.1. Multi-label Classification Techniques

MLC methods are divided in two categories as proposed in [2]: (1) Problem Transformation Methods; and (2) Algorithm Adaptation Methods.

3.1.1. Problem Transformation Methods

It transforms the MLC problem into one or more single-label classification problems. It is an algorithm independent method. Many methods belong tothis category, such as:

- **Binary methods**

- **Binary Relevance (BR):** it is a well-known and the most popular problem transformation method [26]. BR is also known as One-Against-All (OAA). It transforms the multi-label problem into Q-binary problems, by considering the prediction of each label as independent binary classifier. Therefore, BR establishes Q-binary classifiers, one for each label $l \in L$ (whereas: Q = |L|). It transforms original multi-labeled dataset into Q single-label datasets, where each single-label dataset contains all the instances of the original multi-labeled dataset, and trains a classifier on each of these datasets. The instances are labeled positively if they have the existing label, otherwise they are labeled negatively. For the classification of a new instance, it gives the set of labels that are positively predicted by the Q classifiers. Although it is conceptually simple and relatively fast, it is recognized that BR ignores the possible correlations among labels.

❖ **Pair-wise methods**

- **Ranking via Pair-wise Comparison (RPC):** the basic idea of this method is transforming multi-label datasets into q(q-1)/2 binary-label datasets, covering all pairs of labels. (Where q is the number of labels, q = |L|). Each dataset contains the instances of the original multi-labeled dataset that are annotated by at least one of the corresponding labels, but not both. For classifying a new instance, all binary classifiers are invoked. Each classifier votes and predicts one of the two labels. After all classifiers are evaluated, the labels are ranked according to their sum of votes. Then, MLR is used to predict the relevant labels for the intended instance [27].

- **Calibrated Label Ranking (CLR):** it is the extended version of the RPC method [28]. It introduces one additional virtual label V (calibrated label), which is a split point between relevant and irrelevant labels. Thus, CLR solves the MLC problem with the RPC method. Each instance is considered positive if it belongs to the particular label, otherwise it is considered negative for the particular label and positive for the virtual one. The ranking is obtained by summing the votes of all labels; including V. CLR applies both for MLC and MLR tasks.

❖ **Label-combination methods**

These methods remove the limitation of the previous methods by taking into account the correlation and dependencies among labels.

- **Label Power-set (LP):** it is a simple and less-common problem transformation method [2,29]. The idea behind LP is considering each distinct label-set that exists in a multi-labeled dataset as one (single) label to transform the original dataset into a single-label dataset, so any single-label classifier can be applied to it. Given a new instance, the single-label classifier of LP gives the most probable class label, which is actually a set of labels. While the classifier can produce a probability distribution over all class labels, LP can provide the raking task among all. To apply the label ranking, for each label it calculates the sum of probability of class labels that contain it. So, LP can perform MLC and MLR tasks. Although, it takes into account the label correlations, it suffers from the increasing complexity that depends on the large number of distinct label-sets. The number of distinct label-sets is typically smaller, but it is still larger than the total number of labels q (q= |L|), and poses a critical complexity problem, especially for large values of instances and labels.

- **Pruned Set (PS)** [30]: This method follows the same paradigm of LP. But it extends it to resolve its limitations through pruning away the label-sets that are occurring less time than a user-defined threshold. It removes the infrequent label-sets. Then, it replaces these label-sets by the existing disjoint label-sets that are occurring more times than the threshold.

- **Classifier Chains (CC)** [31]: it involves Q-binary classifiers as in a BR method. It resolves the BR limitations, by taking into account the label correlation task. The classifiers are linked along a chain where each classifier deals with the BR problem associated with the label. Each link in the chain is expressed with the 0/1 label associations of all previous links.

❖ **Ensemble methods**

The ensemble methods are developed on top of the common problem transformation and algorithm adaptation methods.

They construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions. They are used for further augment predictive performance and high accuracy results. They aim to aggregate the predictions of several base estimators built with a given learning algorithm.

- **Random k-label sets (RAkEL)** [32]: it constructs an ensemble of LP classifiers. It breaks the large label-sets into m models or subsets, which are associated with random and small-sized k-label-sets. It takes label correlation into account and also avoids LP's problems within the large number of distinct label-sets. Given a new instance, it queries models and finds the average of their decisions per label. Also, it uses the threshold value t to obtain the final prediction. The final decision is positive for a specific label if the average decision is greater than the given threshold t. Thus, this method provides more accuracy of results.

- **Ensembles of Pruned Sets (EPS)** [30]: it combines the PS method in an ensemble scheme. PS is specifically suited to an ensemble due to its fast build times. Also, it counters any over-fitting effects of the pruning process and allows the creation of new label sets at classification time. Applying the ensembles on PS method increases the predictive performance of the algorithm.

- **Ensembles of Classifier Chains (ECC)** [31]: it uses the CC method as a base classifier. It trains m models of CC classifiers C1,C2 ,…,Cm . Each Ck model is trained with a random chain ordering of labels L and a random subset of the datasets D. Each model is likely to be unique and able to predict different label-sets. After that, these predictions are summed by label so that each label receives a number of votes. A threshold value is applied to select the most relevant labels, which form the final predicted multi-label set.

3.1.2. Problem Adaption Methods

It extends and adapts the existing specific learning algorithm to directly handle the multi-label problem. It is an algorithm dependent method. Many methods belong to this category, such as:

- **Multi-Label k Nearest Neighbors (MLKNN)** [25]:it is an extension of the popular k-nearest neighbors (KNN) lazy learning algorithm using a Bayesian approach. It uses the

Maximum A Posteriori principle (MAP) to specify the relevant label-set for the new given instance, based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors. It also has the capability to produce the ranking of the labels.

- **Multi-Label Decision-Tree (ML-DT)** [33]**:** it is an adaptation of the well-known C4.5 algorithm to handle multi-label data. The process is accomplished by allowing multiple labels in the leaves of the tree; the formula for calculating the entropy is modified for solving multi-label problems. The modified entropy sums all the entropies for each individual label. The key property of ML-DT is its computational efficiency:

$$\text{Entropy (D)} = \sum_{i=1}^{q} -p_j \ log_2 \ p_j - \left(1 - \ p_j\right) log_2 \ (1 - \ p_j)$$

Where D is the set of instances in the dataset and p_j is the fraction of instances in D that belongs to the label j.

- **Back-Propagation Multi-Label Learning (BPMLL):**it is a neural network algorithm for multi-label learning. It's derived from the popular basic Back-propagation algorithm. The main modification to the algorithm is the introduction of a new error function that takes multiple labels into account [34].

- **Multi-label Boosting (ADABOOST.MH & ADABOOST.MR)** [35]**:** these are the two extensions of AdaBoost algorithm to handle multi-label data. While AdaBoost.MH is designed to minimize Hamming-loss, AdaBoost.MR is designed to minimize the Ranking-loss and find a hypothesis that ranks the relevant labels at the top.

- **Ranking Support Vector Machine (Rank-SVM)** [24]**:** it is a ranking approach for multi-label learning that is based on SVM. It is used to minimize the Ranking-loss. The main function they use is the average fraction of incorrectly ordered pairs of labels.

- **Multi-label Naïve Bayesian (ML-NB)** [36]**:** it extends the Naïve Bayesian algorithm to adapt it with the multi-label data. It deals with the probabilistic generation among the labels. It uses MAP to specify the more probable labels and assign them to the new given instance.

3.1.3. The Adopted Classification Techniques

We intend to select the most effective and reliable techniques for our experiment. So, looked at the related works that provide a comparison between these algorithms:

1. The authors in [23] compare MLC algorithms: binary relevance (BR), label power-set (LP), random k-label sets (RAKEL) and MLKNN. The RAKEL algorithm is more efficient and gives the best results.

2. The authors in [37] evaluate MLC algorithms RAKEL and MLKNN. Also, RAKEL records the best and effective results.

3. The authors in [38] show that MLKNN provides the best results in almost all analyzed cases.

4. The authors in [39] indicate that, the ECC is the best performance in all measures followed by RAKEL and EPS. The authors observe that, all ensemble methods provide the best results for almost all evaluation metrics.

5. The authors in [40] introduce a survey on the MLC algorithms and states that MLKNN gives better results than other algorithms.

6. The authors in [29 give a detailed description and survey about the MLC algorithms. Then, compare between them by using two different datasets. RAKEL achieves the best results followed by MLKNN. The authors mention that the ensemble methods are the closest algorithms of the best results.

7. The authors in [41] show that the MLKNN performs the best compared to the other algorithms followed by RAKEL algorithm.

From above, we can observe that:

- The algorithm transformation methods: the ensemble methods address the best and most accurate results.

- The algorithm adaptation methods: the MLKNN usually gives higher and best results compared to the other algorithms in the same category.

Therefore, we adopted in our experiment the following MLC techniques:

The Ensemble Methods, from the algorithm transformation category including:

1- Ensemble of Classifier Chains (ECC)
2- Random k-label sets (RAkEL)
3- Ensemble of Pruned Sets (EPS), and
4- Multi-Label k-Nearest Neighbors (MLKNN) from Algorithm Adaptation category.

3.2. Evaluation Metrics

The evaluation of multi-label algorithms requires different measures than those used in single-label classification. Several measures have been proposed for evaluating multi-label classifiers [2,26]. These measures are categorized in three groups: example-based; label-based; and ranking-based metrics. Example-based-measures, evaluate bipartitions over all instances of the evaluation dataset. Label-based measures breakdown the evaluation process into separate evaluations for each label. Furthermore, the ranking-based measures evaluate the ranking of labels with respect to the original multi-labelled dataset. Below, these three types will be described.

However, we need to define some aspects before defining those measures:

- The instances of multi-label dataset (x_i , Y_i), i = 1…. m, where $Y_i \subseteq L$ is the set of true labels and $L = (l_j : j = 1 \ldots q)$ is the set of all labels.

- Given a new instance x_i , the set of labels that are predicted by an MLC algorithm is denoted as Z_i.

- $r_i(l)$ is denoted as the LR method for the label *l*.

3.2.1. Example-Based Measures

- **Hamming Loss:** it evaluates how many times the label of the instance is misclassified, i.e., label which doesn't belong to the instance is predicted or a label belonging to the instance is not predicted. The smaller the value of HL the better the performance:

$$\text{HL} = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \,\Delta\, Z_i|}{|L|}$$

  Where Δ stands for the symmetric difference of two sets, which is the set-theoretic equivalent of the exclusive disjunction (XOR operation) in Boolean logic.

- **Subset Accuracy:** it evaluates the percentage of correctly predicted labels among all predicted and true labels:

$$\text{Subset Accuracy } = \frac{1}{m}\sum_{i=1}^{m} I\,(Z_i = Y_i)$$

  Where I(true) = 1 and I(false) = 0. It is very strict measure where it requires the predicted set of labels to be an exact match of the true set of labels, and ignores predictions that may be almost correct or totally wrong.

The following measurements are:

- **Precision** $= \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|}$
- **Recall** $= \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|}$
- **F_1-Measure**$= \frac{1}{m}\sum_{i=1}^{m}\frac{2\,|Y_i \cap Z_i|}{|Z_i|+|Y_i|}$
- **Accuracy** $= \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i \cup Y_i|}$

3.2.2. Label-Based Measures

These measures are calculated for all labels by using two averaging operations, called macro-averaging and micro-averaging [42]. These operations are usually considered for averaging precision, recall and F-measure. We consider a binary evaluation measures B(tp, tn, fp, fn) which is calculated according to the number of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). The macro and micro-averaged versions of B, can be calculated as follows:

$$\mathbf{B}_{\text{macro}} = \frac{1}{q}\sum_{l=1}^{q} B\,(tp_l, fp_l, tn_l, fn_l)$$

$$\mathbf{B}_{\mathbf{micro}} = \frac{1}{q}\sum_{l=1}^{q} B\left(\sum_{l=1}^{q} tp_l,\ \sum_{l=1}^{q} fp_l, \sum_{l=1}^{q} tn_l, \sum_{l=1}^{q} fn_l\right)$$

3.2.3. Ranking-Based Measures

- **One Error:** it calculates how many times the top-ranked label is not in the set of relevant labels of the instance. The smaller the value of 1-error the better the performance:

$$\text{1-Error} = \frac{1}{m}\sum_{i=1}^{m} \delta(\operatorname{argmin}_{l \epsilon L}\, r_i(l))$$

Where:

$$\delta(l) = 1\ if\ l \notin L, 0\ otherwise$$

- **Coverage**: it evaluates how far we need, to go down the ranked list of labels to cover all the relevant labels of the instance. The smaller the value of coverage the better the performance:

$$\text{Coverage} = \frac{1}{m}\sum_{i=1}^{m} \max_{l \epsilon Y_i} r_i(l) - 1$$

- **Ranking Loss:** it evaluates the number of times that irrelevant labels are ranked above relevant labels. The smaller the value of RL the better the performance**:**

$$RL = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|Y_i||\bar{Y}_i|} \left|\{(l_a, l_b): r_i(l_a) > r_i(l_b), (l_a, l_b) \epsilon\ Y_i x \bar{Y}_i\}\right|$$

where $\bar{Y}_i$ is the complementary set of Y_i with the respect to L.

- **Average Precision:** calculates the average fraction of labels ranked above a particular label $l \epsilon Y_i$ that actually are in Y_i. The bigger the value of AP the better the performance**:**

$$\text{AvgPrec} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|Y_i|} \sum_{l\ \epsilon\ Y_i} \frac{\left|\{l' \epsilon\ Y_i: r_i(l') \leq r_i(l)\}\right|}{r_i(l)}$$

4. EXPERIMENTAL WORK

The LO dataset was created from the ARIADNE repository [15]. It was obtained by using a Web scrapping technique, which is a technique of extracting information from websites. In this experiment, we used the scrapping extension tool attached to Google Chrome browser, called Web Scraper [43]. The dataset should contain a sufficient number of LOs and their related LOM annotations. ARIADNE repository shows the content-related metadata for each browsed LO such as title, description, keywords and rights.

The LO dataset, we have created, contains 658 LO instances, annotated with one or more of 30 labels. These labels correspond to the searched input keywords applied by the learner and to the automatic generation of labels for each LO instance. All labels are related to the computer science domain, such as; computer networks; computer programming; computer graphics; computer security; electronic engineering…. etc. The LOs are described within 3500 features extracted from their LOM annotations In the next subsections, we will explain the approach we followed to automatically assign multiple labels to each LO instance as well as the process of minimizing the size of the metadata features to improve the quality of the classification technique and save the

time. Finally we propose the main statistics of the created multi-labeled dataset.

4.1. Automatic Generation of Metadata (Labels)

Metadata generation is a research field, which has been heavily worked on, in the recent years. Metadata generation method strongly depends on the target metadata types. The focus of this paper is the automatic generation of label metadata and assigning them to the scrapped LO instances. Particularly, this generation is done by keywords metadata. LOs have different keywords. Some of the keywords are different from each other, but their meanings are almost same. Hence, for classification purposes, keywords are categorized, and those categories are used as labels. Label categorization and related keywords are defined and stored in XML file. Then, the parsing function in the java programming language, parses this XML file, and when the LO instance contains any of the listed keywords, the label category of the intended keyword will be assigned to that instance as its additional label. By applying this automatic generation approach, the multiple labels are automatically assigned to each LO [44].

4.2. Dimensionality Reduction (DR)

In machine learning domain, dimensionality reduction (DR) is the process of minimizing and reducing the number of features in the dataset. The motivation for DR is summarized as follows: the reduction of the number of features provides an effective and high accuracy outcomes; the training and classification times are reduced due to the minimization of features' numbers; and removing noisy and irrelevant features which can have an influence on classification and a negative impact on accuracy results.

Dimensionality reduction can be divided into two categories: feature selection and feature extraction [45]. Feature selection is the process of selecting the relevant and high-valued features for the use in dataset classification [23]. Feature extraction is the process that constructs and builds new-derived features out of the original ones; they are intended to be informative and non-redundant.

In this experiment the feature selection approach was used to reduce the features' number. We applied the Gain-Ratio attribute evaluator, from WEKA [46], to select the top valuable features. In the MLC problems, the DR can be executed by invoking one of the multi-label algorithms, as mentioned in (MULAN), a Java Library for Multi-Label Learning, [47]. We performed the attribute evaluation using the LP transformation algorithm.

By applying the DR process, the features' number of the dataset has been reduced from 6166 to 3500 features.

4.3. Dataset Statistics

The multi-labelled dataset has many statistics, which explains the number of labels in the dataset that can influence the performance of the different multi-label methods. These statistics are [26]:

- **Label cardinality:** it is the average number of labels of the instances in dataset:

$$\text{Label-Cardinality} = \frac{1}{m} \sum_{i=1}^{m} |Y_i|$$

- **Label density:** it is the average number of labels of the instances in dataset divided by L (L=

total number of all labels):

$$\text{Label-Density} = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i|}{L}$$

- **Distinct Label-sets:** it provides the number of unique label-sets in the dataset.
- The number of dataset's instances and features, along with features' type: whether they are numeric or nominal.

Table 1. Dataset statistics

			Attributes						
Dataset	Domain	Instances	Before DR	Nominal	Numeric	Labels	Cardinality	Density	Distinct
ARIADNE	Text	658	6166	3500	0	30	2.8586626	0.0952887	299

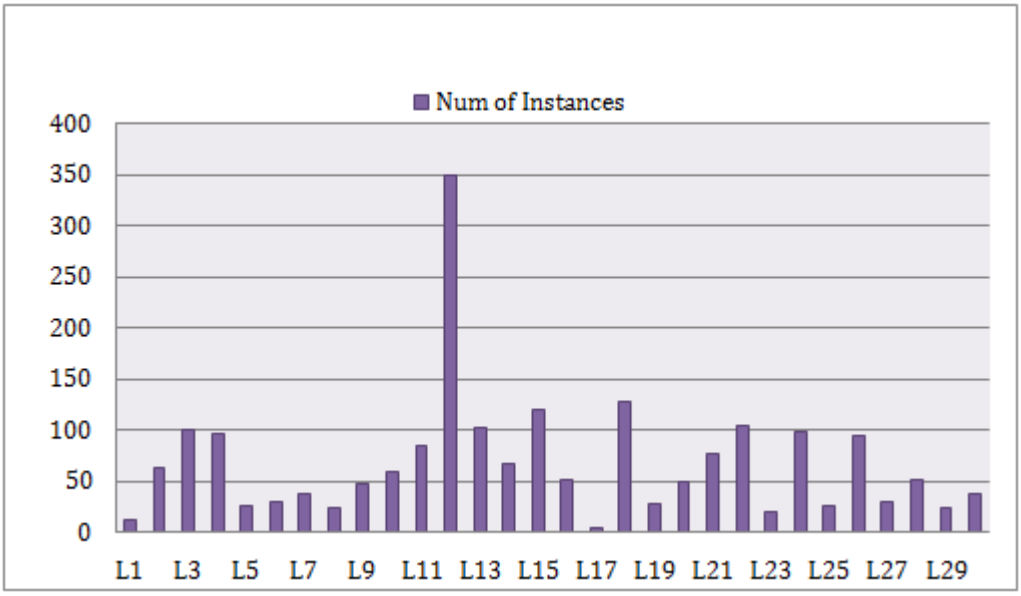


Figure 1. The number of instances per label

5. RESULTS AND DISCUSSION

We have applied the adopted classification techniques from MULAN, for obtaining the predicted results of the dataset. For the experiments, we followed the following three steps of the directive that is available in open-source MULAN system:

1. We loaded the multi-label dataset for training. The dataset is composed of two text files required by MULAN for the specification of a multi-label dataset: an XML file specifying the names of the labels (Ariadne.xml), and an ARFF file specifying the actual data (Ariadne.arff).

2. We created an instance of each learning algorithm that we want to train: ECC, RAKEL, EPS and MLKNN, in order to build a model and obtain predictions.

We trained each classifier using the multi-labeled LO dataset that we loaded from ARIADNE repository. For the empirical evaluation of all adopted algorithms, we used the cross-Validate method of the Evaluator class of MULAN library. Each classifier applied the 4-fold cross validation folds for evaluations to divide the dataset into: training-set and test-set.

The transformation-based algorithms transform MLC problem into one or more single-label problems. So, they accept the single-label classifier (base classifier) as a parameter. In this experiment: the J48 single-label classifier is used as a base classifier for all problem transformation algorithms. J48-classifier is the decision-tree classifier in WEKA Software [48].

Each of the adopted MLC algorithms has it's own parameters, needed to be stated prior to training them.

- **ECC has three parameters:**
  1. The number of models: varied from 30-150 models.
  2. Boolean parameter of using confidence while choosing the subset for dataset: false.
  3. Boolean parameter of using sampling-with-replacement: which means, the instances of the dataset could be selected more than one time at each model: in this paper; it was stated false, each instance could be selected only one time among all models.

- **RAKEL has three parameters:**
  1. The number of models: varied from 50-200 models.
  2. The k-subset size: 3
  3. Threshold value: 0.5

  RAKEL is meta-algorithm, and it can accept any multi-label algorithm as a parameter. It is typically used in conjunction with the LP algorithm. In turn LP is a transformation-based algorithm and it accepts a single-label classifier as a parameter. The J48-classifier, which is the decision-tree algorithm from WEKA, will be used for this purpose.

- **EPS has 6 parameters:**
  1. The percentage of dataset sample at each model: 60%
  2. The number of model: varied from 30-200 models
  3. The threshold value: 0.5
  4. The pruned sets parameter p: 3
  5. The pruned set strategy: Using both strategies: strategy A; and strategy B
  6. The pruned sets parameter b: 3

- **MLKNN has 2 parameters:**
  1. The number of neighbors: varied from 5 to 30 neighbors.
  2. The smooth factor: (always = 1).

5.1. Discussion

The comparison between the four learning algorithms will be evaluated from two points of view:

1. **The Classification point of view:** Table 2 shows the predictive performance results of the four competing MLC algorithms using the evaluation metrics, mentioned above. We noticed that **ECC** dominates the other algorithms in almost all measures, followed by RAKEL, MLKNN and finally EPS. ECC improves the predictive accuracy and can be used to further augment predictive performance.

2. **The Time-Complexity point of view:** In relation to the time issue, we observed that **ECC** is the most time-consuming algorithm, followed by RAKEL algorithm, EPS, and finally MLKNN, which is the fastest algorithm. Table 3 shows the classification time in seconds that was consumed during the process.

From the previous comparison, we could say that **ECC** performs the best and predicts the highest performance. According to the time issue, we have to use special devices, which has a quite enough memory space and a fast processor speed, to do the classification process. In this experiment, we have used our own Laptops to execute the results. Our Laptops have low features compared to more professional devices.

Table 2. Performance results

	ECC	RAKEL	EPS	MLKNN
Example - Based Measures				
Hamming Loss	**0.043918**	0.045595	0.064851	0.062868
Subset Accuracy	**0.326746**	0.323715	0.256790	0.244696
Example-Based Precision	**0.798503**	0.791586	0.794673	0.704624
Example-Based Recall	0.684650	**0.690402**	0.506751	0.511439
Example-Based F Measure	0.702024	**0.703801**	0.566668	0.552293
Example-Based Accuracy	0.617970	**0.618045**	0.478257	0.468276
Label - Based Measures				
Micro-averaged Precision	**0.861877**	0.840858	0.858957	0.830026
Micro-averaged Recall	0.643364	**0.643919**	0.386065	0.428970
Micro-averaged F-Measure	**0.736260**	0.729119	0.531133	0.565065
Macro-averaged Precision	**0.810214**	0.794930	0.437930	0.581787
Macro-averaged Recall	0.578120	**0.590036**	0.277954	0.313773
Macro-averaged F-Measure	0.645468	**0.650262**	0.318188	0.380048
Ranking - Based Measures				
Average Precision	**0.835098**	0.796120	0.715237	0.730319
Coverage	**6.727281**	8.291398	11.435864	8.546322
One-Error	**0.104822**	0.126117	0.136751	0.153427
Ranking Loss	**0.083637**	0.113088	0.170504	0.127692

Table 3. Classification time

Multi-label Classification Algorithm	Time of Classification in Sec
ECC	4012.90 seconds (The slowest)
RAKEL	1650.02 seconds
EPS	181.136 seconds
MLKNN	2.159 seconds (The fastest)

6. CONCLUSIONS AND FUTURE WORK

The services of locating and searching educational contents, specifically LOs, present the core of the development of educational systems. This search area has been active in the recent years. In this paper, we have built an efficient MLC system for classifying the LOs. We have used four effective MLC techniques and compared between them, to notice which classification algorithm is the best for classifying the multi-labelled LOs. The classification was performed on the collection of 658 LO instances and 30 class labels. Therefore this system offers a methodology that illustrates the application of multi-label learning of LOs for classification and ranking tasks. We have concluded that, the **ECC** algorithm was very effective and it was proposed as the best classification algorithm for multi-labelled LOs, followed by RAKEL, MLKNN and finally EPS. From the performance results, it's obvious that the ensemble methods provide the best results for almost all evaluation metrics.

As future work, we intend to: Increase the dataset size, consisting of a very large number of LOs and labels; use the hierarchical MLC approach, which has a great potential in this domain; employ other metadata features to obtain the best classification for LOs; and study the multi-class and multi-instance approaches. They are new studied areas associated with the multi-label learning domain

REFERENCES

[1] Gerard, R. W. (2006). Shaping the Mind: Computers in Education. In R. C. Atkinson and H. A. Wilson (eds.), Computer-Assisted Instruction: A Book of Readings. Orlando, Fla.: Academic Press, Health Education Assets Library.

[2] Tsoumakas, G., and Katakis, I. (2007). Multi-label classification an overview. International Journal of Data Warehousing and Mining. 3(3), 1-13

[3] Zhu, S.; Ji, X.; Xu, W.; and Gong., Y. (2005). Multi-labelled classification using maximum entropy method. In Proceedings SIGIR, 1-8.

[4] Barutcuoglu, Z.; Schapire, R.; and Troyanskaya, O. (2006). Hierarchical multi-label prediction of gene function. Bioinformatics 22, 880–836.

[5] Gerard, R.W. (1967). Shaping the mind: Computers in education. In National Academy of Sciences, Applied Science and Technological Progress. 207-228

[6] Lehman, R. (2007). Learning Object Repositories. New Directions for adult and continuing education, Spring. Wiley Periodicals, Inc, 113, 57-65. Retrieved from www.interscience.wiley.com

[7] Semmens, P. N. (2004). The potential for learning objects to support flexible learning in higher education. IEEE Computer Society Technical Committee on Learning Technology newsletter 6(2), 1-5. Retrieved from http://www.ieeetclt.org/issues/april2004/learn_tech_april2004.pdf .

[8] Polsani, P.(2003). Use and abuse of reusable learning objects. Journal of Digital Information, 3(4), 1-10. Retrieved from http://www.info2.uqam.ca/~nkambou_r/DIC9340/seances/seance10et12/Standards%20et%20LO/http___jodi.ecs.soton.ac.pdf.

[9] IEEE Learning Technology Standards Committee (LTSC). (2013, October 6). Retrieved from http://ieee-sa.centraldesktop.com/ltsc/

[10] DCMI Home: Dublin Core® Metadata Initiative (DCMI). Retrieved from http://dublincore.org/

[11] IEEE. (2006). IMS Meta-data best practice guide for IEEE1484.12.1-2002 Standard for Learning Object Metadata. Retrieved from http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html .

[12] The ADL SCORM specification v 1.2. (2003). ADL – Advanced Distributed Learning Initiative, SCORM –Shareable Content Object Reference Model , 1(2), 9. Retrieved from http://www.adlnet.org

[13] CanCore: Homepage. Retrieved from http://cancore.athabascau.ca/en/index.html

[14] MERLOT II - Home. Retrieved from https://www.merlot.org/merlot/index.htm

[15] Ariadne Foundation. Retrieved from http://www.ariadne-eu.org/

[16] NSDL Library. Retrieved from https://nsdl.oercommons.org/

[17] EHSL - HEAL Collection. Retrieved from http://library.med.utah.edu/heal/

[18] Educational Network Australia. Retrieved from http://www.network-ed.com.au/

[19] Comp R. Cerri, R. R. Silva, and A. C. Carvalho,. “Comparing Methods for Multilabel Classification of Proteins Using Machine Learning Techniques”, BSB 2009, LNCS 5676, 109-120, 2009.

[20] Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. Pattern Recognition, 37(3), 1757–1771. Retrieved from https://www.rose-hulman.edu/~boutell/publications/boutell04PRmultilabel.pdf

[21] Qi, G., Hua, X., Rui, Y., Tang, J., Mei, T., & Zhang, H. (2007). Correlative multi-label video annotation. In Proceedings of the 15th international conference on Multimedia (MULTIMEDIA '07). ACM, New York, 17-26.

[22] Diplaris, S., Tsoumakas, G., Mitkas, P. A., & Vlahavas, I. P. (2005). Protein Classification with Multiple Algorithms. Springer, 10th Panhellenic Conference on Informatics, PCI 2005, Volas, Greece, November 11-13, 2005. Proceedings,3746, 448-456.

[23] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008). Multi-Label Classification of Music into Emotions. In: Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, 6.

[24] Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. In: Advances inNeural Information Processing Systems, 1-7.

[25] Zhang, M., & Zhou, Z. (2007). ML-KNN: A lazy learning approach to multi-label learning. ScienceDirect, Pattern Recognition, 40(7), 2038-2048.

[26] Tsoumakas, G., Katakis, I., & Vlahavas, I. P. (2010). Mining Multi-label Data,1-20.

[27] Hüllermeier, E., Fürnkranz, J., Cheng, W., & Brinker, K. (2008). Label ranking by learning pairwise preferences. Artificial Intelligence, 34(5), 1897–1916.

[28] Fürnkranz, J., Hüllermeier, E., & Brinker, K. (2008). Multi-label classification via calibrated label ranking. ACM- Digital Library, Machine Learning, 73(2), 133 - 153.

[29] Sorower, M. S. (2010). A Literature Survey on Algorithms for Multi-label Learning, 25.

[30] Read, J., Pfahringer, B., & Holmes, G. (2008). Multi-label Classification Using Ensembles of Pruned Sets. Proc 8th IEEE International Conference on Data Mining, Pisa, Italy, 995-1000.

[31] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. Springer,Machine Learning, 5782, 254-269.

[32] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Random k-Labelsets for Multi-Label Classification. Knowledge and Data Engineering, IEEE Transactions, 23(7), 1079 - 1089.

[33] Clare, A., & King, R. D. (2001). Knowledge Discovery in Multi-label Phenotype Data.in: Proceedings of the 5th European Conference on PKDD, 42-53.

[34] Zhang, M., & Zhou, Z. (2006). Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering, 18(10), 1338-1351.

[35] Schapire, R. E., & Singer, Y. (2000). BoosTexter: a boosting-based system for text categorization. Machine Learning, 39(2/3), 135-168.

[36] Wei, Z., Zhang, H., Zhang, Z., Li, W., & Miao, D. (2011). A Naive Bayesian Multi-label Classification Algorithm With Application to Visualize Text Search Results. International Journal of Advanced Intelligence, 3(2), 173-188.

[37] Batista, V., Pintado, F. P., Gil, A. B., Rodríguez, S., & Moreno, M. (2011). A System for Multi-label Classification of Learning Objects. Springer, Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011 Advances in Intelligent and Soft Computing, 87, 523-531.

[38] Santos, A. M., P Canuto, A. M., & Neto, A. F. (2011). A Comparative Analysis of Classification Methods to Multi-label Tasks in Different Application Domains.International Journal of Computer Information Systems and Industrial Management Applications, 3, 218-227.

[39] El Kafrawy, P., Mausad, A., & Esmail, H. (2015). Experimental Comparison of Methods for Multi-Label Classification in Different Application Domains. International Journal of Computer Applications, 114(19), 1-9.

[40] Prajapati, P., Thakkar, A., & Ganatra, A. (2012). A Survey and Current Research Challenges in Multi-Label Classification Methods. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 248-252.

[41] Tawiah, C. A., & Sheng, V. S. (2013). Empirical Comparison of Multi-Label Classification Algorithms. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 1645-1646.

[42] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1, 67–88.

[43] Web Scraper - Chrome Web Store. Retrieved from https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn?hl=en

[44] Meyer, M., Rensing, C., & Steinmetz, R. (2007). Categorizing Learning Objects Based On Wikipedia as Substitute Corpus. Proceedings of the First International Workshop on Learning Object Discovery & Exchange, 64-71.

[45] Sun, L., Ji, S., & Ye, J. (2014). Multi-label dimensionality reduction (1st ed.). USA: Chapman and Hall/CRC.

[46] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/

[47] Mulan: A Java library for multi-label learning. Retrieved from http://mulan.sourceforge.net/index.html

[48] J48-Decision Tree Classifier- WEKA. Retrieved from http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

IMPROVEMENT OF A METHOD BASED ON HIDDEN MARKOV MODEL FOR CLUSTERING WEB USERS

Sadegh Khanpour[1] and Omid sojoodi[2]

[1]Faculty of Electrical, Computer and IT Engineering,
Qazvin Azad University, Qazvin, Iran
sadeghkhanpour@gmail.com
[2]Faculty of Electrical, Computer and IT Engineering,
Qazvin Azad University, Qazvin, Iran
O_sojoodi@qiau.ac.ir

ABSTRACT

Nowadays the determination of the dynamics of sequential data, such as marketing, finance, social sciences or web research has receives much attention from researchers and scholars. Clustering of such data by nature is always a more challenging task. This paper investigates the applications of different Markov models in web mining and improves a developed method for clustering web users, using hidden Markov models. In the first step, the categorical sequences are transformed into a probabilistic space by hidden Markov model. Then, in the second step, hierarchical clustering, the performance of clustering process is evaluated with various distances criteria. Furthermore this paper shows implementation of the proposed improvements with symmetric distance measure as Total-Variance and Mahalanobis compared with the previous use of the proposed method (such as Kullback–Leibler) on the well-known Microsoft dataset with website user search patterns is more clearly result in separate clusters.

KEYWORDS

Hidden Markov Model, distance metric, agglomerative clustering, categorical time series sequence, probability model

1. INTRODUCTION

Determining the dynamics in a Sequential data has become a critical step in many research fields. Current researches on data mining methods for dealing with big data in clustering of sequential data has recently aroused great interest[1] For example, discovering patterns in web navigation, similar to web mining, has become an important subject[3]. In this regard, [2] it was illustrated that traditional data mining approaches might be unsuitable for pattern discovery of websites users. Therefore, a large number of algorithms have been proposed for clustering web usage patterns. For example, the approaches that use K-means algorithm with KL distance metric as an alternative of Euclidian-metric distance [4], are resulted in the development of hierarchical model

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 125–136, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60211

based algorithms for web transactions clustering [5] and model-based approaches based on Markov models [6].

Along with data mining techniques, the use of probabilistic models such as Markov chain model [9], is useful in classification of web pages and generation of similarity and relation between different web sites. In order to conduct web mining, information from various sources such as web server access log, proxy server log, log browser, user profiles, data registration and meeting user transactions, cookies, bookmarks data , mouse clicks, surveys and other data can be collected as a result of an interaction.

In [10] it was investigated that the Markov models in web mining can be used to predict the user's next action; for example, using Markov models, social networks can predict future visits of users. Social networks can be mapped as a Markov chain; also using hidden Markov models with support vector machine classification methods, predicting sports, weather and social activities on Twitter was possible.

HMM is a machine learning algorithm used for pattern recognition in various applications (e.g. speech recognition, text and movement). The algorithm consists of two random processes. Hidden processes are not visible directly but indirectly can be deduced throughout the random process that produces a sequence of observations. Statistical methods such as Markov models can be employed to explore the behavior of transient (temporary) web data.

In section 2 we review previous researches on the application of different types of Markov models in various fields of web mining. Section 3 introduces the issue and the constraints involved in solving them, using previous methods. Section 4 describes the process of modeling and hierarchical clustering problem using different distance-metric criteria. Section 5 describes the standardized data set that contains records of web users' browsing history by introducing and applying the proposed method. Section 6 compares the quality of clustering, using different distance measures, and reports findings and results. Section 7 and 8 present future work and references list respectively.

2. REVIEW OF RESEARCH ON THE APPLICATION OF MARKOV MODELS IN WEB MINING

Traditional hierarchical clustering algorithms commonly used in clustering are somewhat impractical because it requires more storage and computation when the number of observations is large. K-means algorithm is considered as one of the most widely used algorithms in web mining. For the clustering of web users and user sessions, modeling studies based on Boolean (met / not met) or based on the frequency (number of each page visits) were adapted in web application. In other studies exploring the general sequence, sequence pattern mining techniques are used in order to reduce the computational complexity and produce significant clusters (meaningful). Using statistical models such as Markov models, in particular in the clustering process and display data encoding, is a more efficient way, so that the review of the current paper, Markov chain models role is well-appreciated with capabilities in three areas of web mining (usage, content, and structure mining). C. Xu et al., [13] proposed a hidden semi Markov model in web usage mining that the page sequence {page1, page2, … } can be described as a Markov chain; the HTTP requests $\{r_1, r_2, \dots, r_n\}$ or interval time $\{O_1, O_2, \dots, O_{n-1}\}$ between adjacent requests can be treated as the observations of Markov chain; and each Markov state can output multiple

observations continuously. Obviously, the user click behaviour conforms to hidden semi-Markov model (HsMM). Their methods were used state selection algorithm based on K-means clustering on backbone of a state China Telecom data set. Luca De Angelis et al., [14] proposed an extended hidden Markov model and time series in web usage mining. They mine categorical sequences from data using a hybrid clustering method that observation was sequences (variable length) of change transition between states and hidden states generated dynamics by time series, training algorithm was EM on well-known Microsoft dataset with website users search patterns. Sungjune Park et al., [15] proposed Markov chain in web usage mining. Parameters were page categories as state that each Markov chain represents the behaviour of a specific subgroup and categorize page with K-means & Kmeans + Fuzy ART methods, training algorithm was EM on Information Server (IIS) logs for **msnbc.com** data set for the entire day of September, 28, 1999. Yu-Shiang Hung et al., [16] proposed combined Markov model with ART2-enhance in web usage mining, each Markov chain represents the behaviour of a specific subgroup and training algorithm was EM on all of march, 2012, 3391 sessions of 157 elders data set were identified for analysis. Yi Xie et al., [17] proposed a Large-scale hidden semi Markov model for web security in web usage mining. Parameters were pages of web site as hidden state that states transition was the structure of web page links and inner requests of pages as observation. Training algorithm was unsupervised extended re-estimation [21,22] and entropy clustering on anomaly detection in behaviours of user navigation data set.

3. PROBLEM DEFINITION

As seen in the example of table 1, sequences may be restrictions on the issue of calculating the distance (similarity / dissimilarity) between them exist.

Table 1. Observed sequences A and B.

Time	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Sequence A	1	1	1	2	2	2	2	2	1	1	1	2	1	2	1	2
Sequence B	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2

Table 2. observed transitions between states.

Sequence A		
States	1	2
1	4	4
2	3	4

Sequence B		
States	1	2
1	4	4
2	3	4

Table 2 contains two categorical sequences of A & B with state- space {1 and 2} and length of 16. According to the table 1, the sequences are different. In fact, from a Markov Chain perspective these two sequences are identical as its sufficient statistics are the same.As it is shown in table 2, the starting condition and state of the state transition matrix are identical in observations. Therefore, based on these data, applying any distance-metric between A and B would be null; in other words, the sequences would be identified as identical which always results in the belonging to the same cluster.

The aim of this paper is to improve method that based on hidden Markov model in [14] clustering the search pattern of web users. The proposed improvements on method of this paper are a

combination of model-based clustering approach, which is built through the development of one HMM. In particular, the process of clustering with correlated data is observed, developed in time series and categorical data set is transformed to a probabilistic space in a way that the symmetric distance-metrics as Kullback–Leibler, Total-Variance and Mahalanobis could be applied on it.

The BCD algorithm clusters the time series based on the distances between the matrixes of observed transitions between states; while the procedure of clustering in this paper rests on the distances between posterior probabilistic of hidden states resulted from HMM after learning phase (estimated using Baum Welch algorithm[24]). Additionally, the estimation from HMM panel the similar hidden part of each of the time series lets the last probability of each sequence be comparable.

4. PROBLEM MODELING AND CLUSTERING

This section introduces a flexible method for categorical clustering of times series and clustering procedures based on model and hierarchical.

4.1. Definintion, Concepts and Notation

Y is a sample of n objects from time series sequences and each of its, with variable length (from 1 to T_i) that t={1, 2 , … , T_i}, denoted by Y = ($Y_1, Y_2, \dots, Y_n$) subject to $Y_i = (Y_{i\,1}, Y_{i\,2}, \dots, Y_{i\,T_i})$ and Z = ($Z_1, Z_2, \dots, Z_n$) explains different hidden states of Markov. φ Explains the set of parameters which is f(y_i; ϕ) the probability density function for object i with the particular parameter φ. The logarithm of maximum likelihood (ML) function of data for the set of parameters is $l(\varphi; y) = \sum_{i=1}^{n} \log f(y_{i;}\, \varphi)$.

$$f(Y, Z \mid \phi) = ? \tag{1}$$

$$\text{data set is } Y = \{ Y_1, Y_2, \dots, Y_n \} \tag{2}$$

$$Z = \{ Z_1, Z_2, \dots, Z_n \} \tag{3}$$

$$| Y_i | = T_i \;, Y_i = \{ Y_{i\,1} = A, Y_{i\,2} = B, \dots, Y_{i\,T_i} = \cdots \} \;, Y_{i\,j} \in \{ 1, \dots, M \} \tag{4}$$

$$| Z_i | = T_i \;, \; Z_i = \{ Z_{i\,1}, Z_{i\,2}, \dots, Z_{i\,T_i} \} \;, Z_{i\,j} \in \{ 1, \dots, K \} \tag{5}$$

4.2. Step 1 : HMM panel

The first step, provides a model-based approach through the development of the concept of hidden Markov model. It assumes that time observations (time series) Y_t is dependent on the hidden random process of Z_t which is defined with K states. The relation between data series is observed and the hidden process for object i is shown in picture 2.

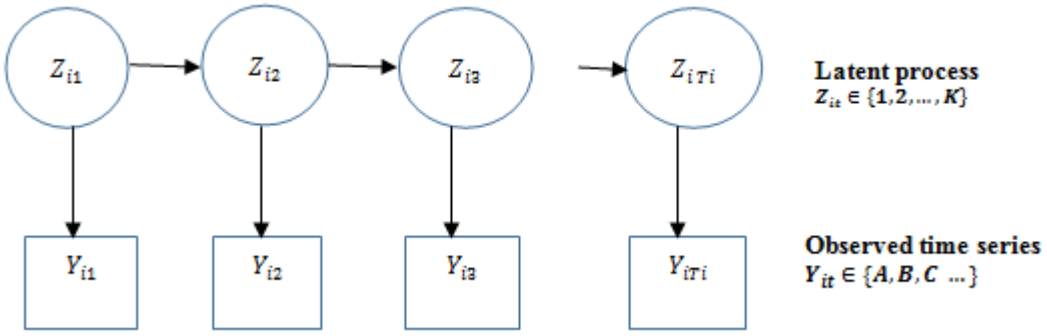


Fig. 2. Graphical representation of the model

The above graph shows the serial dependency in the observations which are completely affected by Z_t process; thus, HMM with the defined variable from different states of K for each time observation, HMM estimates the hidden variable for the entire T_i. For the set of ϕ parameters, HMM panel is determined for object i as follow:

$$\text{HMM } f(Y_i, Z_i|\phi) = f(y_{i1} | z_{i1}).\ f(z_{i1}) \prod_{t=2}^{T_i} f(y_{it} | z_{it}) f(z_{it} | z_{i(t-1)}) \quad (6)$$

$$\text{and } f(Y_i | \phi) = \sum_{z_i} f(Y_i, Z_i|\phi) \quad (7)$$

$$f(Y, Z | \phi) = \prod_{i=1}^{n} f(y_i, z_i | \phi) \quad (8)$$

$$\text{all i} : \lambda_i \equiv f(z_i = k) \ ; \ \sum_{i=1}^{n} \lambda_k = 1 \quad (9)$$

$$A \text{ (latent state transition matrix)} \equiv \pi_{wk} \equiv f(z_{it} = k \mid z_{it-1} = w); \sum_{k=1}^{K} \pi_{wk} = 1 \quad (10)$$

$$B \text{ (observe state transition matrix)} \equiv B_{kj} \equiv f(Y_{it} = j \mid z_{it} = k) \ ; \sum_{j=1}^{M} B_{kj} = 1 \quad (11)$$

Now, due to high number of parameters estimation of $\hat{\varphi} = argmax_{\varphi} l(Y | \phi)$ using ML method would be very difficult. Therefore, approximation method of EM, which is Baum Welch for HMM, will be used.

After finishing the learning phase of HMM, the set of $\hat{\varphi}$ ($\widehat{\lambda_k}$ ،$\widehat{\pi_{wk}}$ و $\widehat{B_{kj}}$) parameters will be estimated and they let us calculate the posterior probabilities. The $\widehat{u_{ik}}(t) = f(\ i \in k \text{ at time } t \ | y_i, \hat{\varphi})$ which is the probabilities of an observation and is estimated in a hidden states in time *t* condition on time series observations and estimated parameters. In that paper, transformation of the main dataset to $\widehat{u_{ik}}(t)$ posterior probabilities is suggested. This transformation provides two advantages: 1. Posterior probabilities contain exclusive information of each time series and this could be used simply in clustering. 2. They can be compared; for instance, for $k = 1, 2, \dots, K$ and $\widehat{u_{ik}}(t)=1$, $i = 1, 2, \dots, n$. The most important analytical step of this paper is the determination of hidden conditions (K) of Markov process as in [14]. The value of K is identified 12 in the implementation of web mining on msn web data set(available on kdd.ics.uci.edu).

4.3. Step 2: Hirachical Agglomerative Clustering

The second step is to extract the information provided by the probabilities obtained in the first step to determine clusters of sequences characterized by similar dynamic patterns. Specifically, if two time series with posterior probability are assigned (allocated) to the same hidden status, then they should be in the same cluster. Natural chance to calculate the distance between two probability distributions will be KL.

All t,i : $f(z_{it(1..T_i)} = k \mid Y_{i(1..T_i)})$ (12)

Table 3. Mapping i'th sequence to form the posterior probabilities

	1	2	…	t	…	
1						
2				…		
K			…		…	
...				…		
K						

4.4. Distance Metric Improvement in Clustering Phase

Instead of KL similarity measure, other criteria those are applicable to the data with statistical distribution, such as Total Variance and Mahalanobis, can be used.

4.4.1. TV Distance Metric [11]

$q = \widehat{u_{jk}}$, $p = \widehat{u_{ik}}$ (and the opposite) $D_{TV}(p \parallel q) = ½ \sum_{i=1}^{M} | p(i) - q(i)|$ (13)

4.4.2 Mahalanobis Distance Metric [12]

This distance criteria for an observation $Y = (y_1, y_2, \dots, y_N)^T$ from a group of observations with the average observation $\mu = (\mu_1, \mu_2, \dots, \mu_N)^T$ and covariance matrix S is defined as follows:

$$D_M(Y_i, Y_j) = \sqrt{(Y_i - \mu)^T S^{-1} (Y_j - \mu)} \quad Y = (y_1, y_2, \dots, y_N)^T \tag{14}$$

The steps of implementation of Mahalanobis distance measure by using principal component analysis PCA [11]:

- Calculation of the total length of time series sequence data: $T_{total} = \sum_{i=1}^{n} T_i$

- With concatenation of probability matrix for each of the sequences, the total matrix has the dimension of the $T_{total} * K$ is formed (K number of hidden states).

- execution of PCA feature extraction algorithm on the previous step matrix (Without dimension reduction)

  - eigenvectors matrix (each column is a special-vector in descending order)

  - the transformed total matrix of the PCA space

  - variance in the directions of the eigenvectors

- To be in line with the amount of variance should be divided every direction, the standard deviation of direction.

5. WEB USAGE MINING

In this section, using the proposed method, msnbc.com (available in kdd.ics.uci.edu) dataset have been analysed with many distance measurements. The application aims to analyze the sequence of pages requested by the user search patterns visits to determine the clusters identified through different websites and search behavior on the Web. This data set is also used by other researchers [18, 19].

5.1. Data Set Description

Dataset includes a record number of 989,818 (each record a sequence of different pages a user visits) which is recorded by Microsoft MSN for websites visitors during a full day (28 September 1999). The variety of the visited web pages are 1 to 17:

(1) frontpage, (2) news, (3) tech, (4) local, (5) opinion, (6) on-air, (7) misc, (8) weather, (9) health, (10) living, (11) business, (12) sports, (13) summary, (14) bbs (bulletin board service), (15) travel, (16) msn-news, and (17) msn-sports.

Z considered hidden states of model of HMM, Y is Set of sequences of variable length, each of the constituent elements of each sequence in the moment t (t is 1 to T_i) have one of the 17 categories may be. The analysis and application of the proposed method in a random sample of about one percent of the data that has at least two pages have (have at least one transition) sample size n = 6244 are considered.

5.2. Step 1: HMM panel

In the first step of HMM panel, for reaching the optimum number of hidden state, K with values of 1 to 15 is analyzed and in order to avoid local maxima, the algorithm is run for 100 times. The best K (the number of hidden states for each item in the sequence) was determined 12.

5.3. Step 2: Hierarchical Agglomerative Clustering with Different Distance Metrics

Now we compute the distance between the two sequences with different distance metrics , then in the step 2, we start to cluster the obtained distance matrix with the usage of hierarchical agglomerative clustering algorithm with different distance method between clusters such as complete linkage method. As can be seen in the following figures (3, 4, 5, 6 and 7) according to the dendrograms' using of the different distance metrics, we are cutting the tree with the biggest cut, one of cluster that consists a large number of users, include the users who prefer short meetings with specific topics such as news headlines a few clicks, weather and sports search. Thus we could label this cluster as specialists' users. The second cluster represent generalist users such as web users have longer sessions characterized by longer sequences of clicks and prefer different topics on the website.

6. EVALUATION AND RESULTS

Applying KL, TV and Mahalanobis distance criteria with different distance methods from the figures of part 5.3 it is concluded that the implementation of the second part of the algorithm with the criteria of TV-ward, TV-complete, Mahalanobis-complete (with respect to the maximum distance between levels of clustering) respectively have more clearly result in separate clusters than distance measure (KL) utilized in the referenced paper in the clustering level.

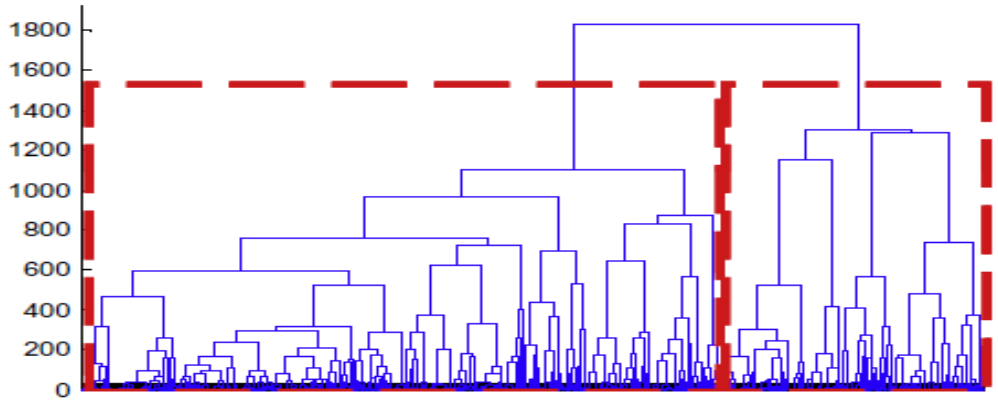


Fig. 3. Dendrogram -using KL with complete method (Basis for comparison- *fourth choice*)

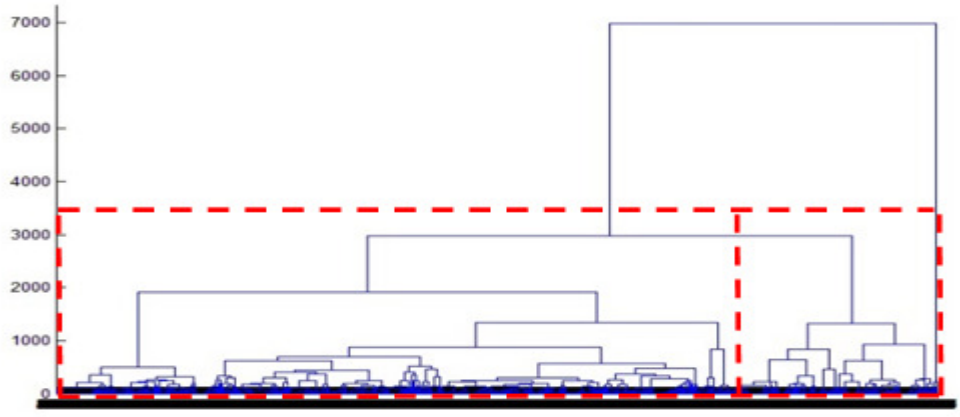


Fig. 4. Dendrogram -using TV with ward method (The best of the three criteria-*first choice*)

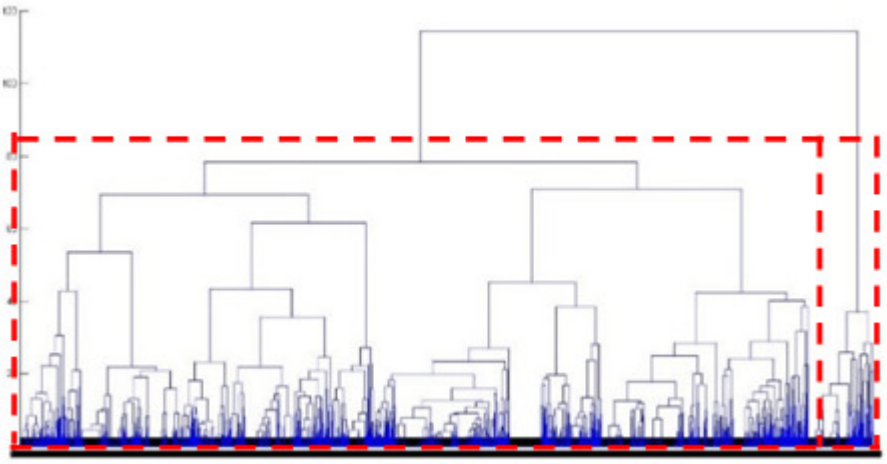


Fig. 5. Dendrogram -using TV with complete method(*second choice*)

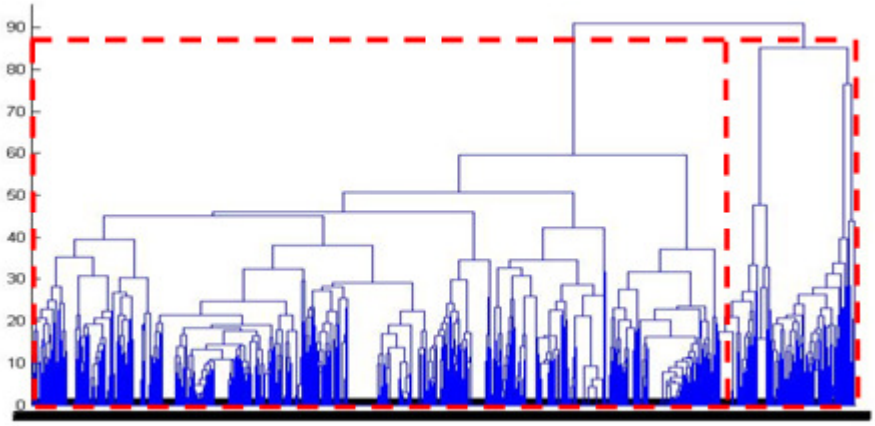


Fig. 6. Dendrogram -using Mahalanobis with complete method(better than KL – *third choice*)

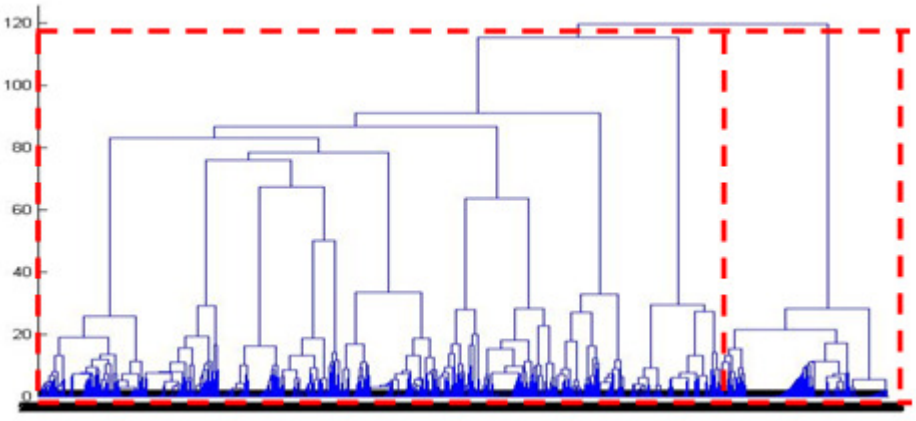


Fig. 7. Dendrogram -using Mahalanobis with complete method (*worse than KL and thired choice*)

7. CONCLUSIONS

In this paper we have improved a method based on hidden Markov model for clustering web users with different symmetric distance measures. Among the measures discussed, as Total-Variance and Mahalanobis compared with the previous work that was used of the proposed method (such as Kullback–Leibler) are better results and more clearly separated clusters as test results show up.

8. FUTURE WORKS

In order to development of the current work, we can do the next phase HMM training and clustering, we can predict the next sequence that improvement attractive website for website designers, and predict future patterns of search engine users may be fruitful. The clustering is done using a combination of improved methodology presented in this paper for other applications, such as browsing behaviour anomaly detection website user [17] and also on the results of other data collection including the development of the proposed method in this paper.

REFERENCES

[1] Ananthanarayana, V. S., Murty, M. N., & Subramanian, D. K. (2001). Efficient clustering of large data set. Pattern Recognition, 34, 2561–2563.

[2] Spiliopoulou, M., & Pohle, C. (2001). Data mining for measuring and improving the success of Web sites. Data Mining and Knowledge Discovery, 5(1–2), 85–114.

[3] Vakali, A., Pokorny, J., & Dalamagas, T. (2004). An overview of Web data clustering practices (pp. 597–606). Berlin: Springer.

[4] Petridou, S. G., Koutsonikola, V. A., Vakali, A. I., & Papadimitriou, G. I. (2006). A divergence oriented approach for Web users clustering. In M. e. a. Gavrilova (Ed.), ICCSA 2006. LNCS 3981 (pp. 1229–1238). Heidelberg: Springer.

[5] Yang, Y., & Padmanabhan, B. (2011). Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies. European Journal of Operational Research, 215(3), 679–687.

[6] Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. Data Mining and Knowledge Discovery, 7(4), 399–424.

[7] Ramoni, M., Sebastiani, P., & Cohen, P. (2002). Bayesian clustering by dynamics. Machine Learning, 47(1), 91–121.

[8] Ramoni, M., Sebastiani, P., & Kohane, I. (2002). Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences of the United States of America, 99, 9121–9126.

[9] V.Chitraa, Dr. Antony Selvdoss Davamani, A Survey on Preprocessing Methods for Web Usage Data Information Retrieval System, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010

[10] Dias, J. G., Cortinhal, M. J. (2008). The skm algorithm: A k-means algorithm for clustering sequential data. In Geffner, H., Prada, R., Alexandre, I. M., David, N.(Eds.), Proceedings of the advances in artificial intelligence – Iberamia. Lecture notes in computer science (vol. 5290, pp. 173–182).

[11] Sergios Theodoridis, Pattern Recognition, 4th edition, Copyright © 2009, Elsevier Inc. All rights reserved.

[12] Bishop, Pattern Recognition and machine learning, 2end edition, 2006 Springer Science, Business Media, LLC

[13] C. Xuchuan, C. Dua, G.F. Zhao, S. Yu, A novel model for user clicks identification based on hidden semi-Markov, Journal of Network and Computer Applications 36 (2013) 791–798, ELSEVIER.

[14] Luca De Angelis, José G. Dias,Mining categorical sequences from data using a hybrid clustering method, European Journal of Operational Research xxx (2013), ELSEVIER.

[15] Sungjune Park, Nallan C. Suresh, Bong-Keun Jeong, Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm, Data & Knowledge Engineering 65 (2008) 512–543, ELSEVIER.

[16] Yu-Shiang Hung, Kuei-Ling B. Chen, Chi-Ta Yang, Guang-Feng Deng, Web usage mining for analysing elder self-care behavior patterns, Expert Systems with Applications 40 (2013) 775–783, ELSEVIER.

[17] Yi Xieyicn and Shun-Zheng , A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 17, NO. 1, FEBRUARY 2009.

[18] Dias, J. G., & Vermunt, J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation. Journal of Retailing and Consumer Services, 14(6), 359–368.

[19] Ramos, S., Vermunt, J., & Dias, J. (2011). When markets fall down: Are emerging markets all the same? International Journal of Finance and Economics, 16(4), 324–338.

[20] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.

[21] M. Kantardzic, WEB MINING A ROADMAP, Methods And Algorithm. New York: IEEE Press, 2009.

[22] X.Yi and Y. Shunzheng, "A dynamic anomaly detection model for web user behavior based on HsMM," in Proc. 10th Int. Conf. Computer Supported Cooperative Work in Design (CSCWD 2006), Nanjing, China, May 2006, vol. 2, pp. 811 816

[23] Bamshad Mobasher, 'Data Mining for Web Personalization' , School of Computer Science, Telecommunication, and Information Systems DePaul University, Chicago, Illinois, USA

[24] Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. IEEE Information Theory Society Newsletter, 53:1(4), 10–13.

AUTHORS

First Author Sadegh Khanpour is PhD student of Qazvin Islamic Azad University, as well as the director of the data mining research and development team from one of the chain stores in Iran.

Second Author Omid Sojoodi holds a PhD in Artificial Intelligence from the UPM University of Malaysia, and faculty member of Islamic Azad University of Qazvin.

GRASP APPROACH TO RCPSP WITH MIN-MAX ROBUSTNESS OBJECTIVE

Hayet Mogaadi and Besma Fayech Chaar

National Engineering School of Tunis, El Manar University, Tunisia
mogaadi_h@yahoo.fr
besma.fayechchaar@insat.rnu.tn

ABSTRACT

This paper deals with the Resource-Constrained Project scheduling Problem (RCPSP) under activity duration uncertainty. Based on scenarios, the object is to minimize the worst-case performance among a set of initial scenarios which is referred to as the min-max robustness objective. Due to the complexity of the tackled problem, we propose the application of the GRASP method which is qualified as a simple and effective multi-start metaheuristic. The proposed approach incorporates an adaptive greedy function based on priority rules to construct new solutions, and a local search with a forward-backward heuristic in the improvement phase. Two different benchmark data sets are investigated, the Patterson set and the PSPLIB J30 set. Comparative results show that the proposed enhanced GRASP outperforms the basic procedure in robustness optimization.

KEYWORDS

RCPSP, uncertainty, Robustness, scenario, GRASP, intensification

1. INTRODUCTION

The Resource-Constrained Project Scheduling Problem (RCPSP) is a well-known project scheduling problem that consists in scheduling a set of activities over resources with limited capacities subject to precedence and resource constraints while optimizing several objectives. The most common objective is to minimize the project duration (so called makespan and denoted by Cmax). As a generalization of a majority of the classical scheduling problems, the RCPSP was classified as NP-hard [1]. This problem was widely studied in the literature; heuristics and metaheuristics were successfully applied in this context such as Genetic Algorithms, Sampling methods, based local search methods, Simulated Annealing, etc. Efficient surveys are given in the following references [2, 3].

Nevertheless, the project scheduling process is really subject to unexpected events that may be related to activities or resources leading, in the most of cases, to schedule disruptions. In the last few decades, the researchers 'efforts were focused in managing uncertainty in project scheduling to avoid the schedule disruption and the performance degradation when perturbations occur.

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 137–146, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60212

One of the basic approaches to deal with uncertainty [4] is the robust approach having the object to find a schedule that remains with a highest quality across a set of scenarios. A scenario represents a problem realization which is founded by matching fixed values to uncertain problem parameters. Inspired from the decision analysis, the scenario-based approach is a simple and effective way to model uncertainty. With the absolute robustness objective, referred to as the min-max objective, the aim is to minimize the maximum performance degradation among all scenarios. However, the regret robustness objective is to minimize the maximum deviation of solutions from optimality across all scenarios.

In the literature [5], Kouvelis and Yu have investigated the cited robustness objectives for different combinatorial optimization problems. Although robust scheduling problems are more blinded to reality, solution methods for robust RCPSP are not exhaustive. In [6], Al Fawzen and Haouari have proposed a bi-objective model for RCPSP with the minimization of the makespan and the maximization of the robustness. The problem was solved by a tabu search heuristic. Chtourou and al. [7] have studied various robustness measures based on priority rules when activity durations vary. The work of Artigues and Leus [8] deals with RCPSP under activity uncertainty. Based on PLNE, the authors proposed a scenario-based bi-level problem formulation that minimizes the absolute and relative regret robustness. In this model, a solution depends on priority rule, also called scheduling policy. The authors have applied, in first, exact method which has taken excessive computational time considering medium sized instances. So they were directed towards heuristic procedures. In addition, the Genetic algorithm was simply adapted to robust optimization problems, such as the one machine problem [9] and the robust RCPSP [10].

Heuristics and metaheuristics are also approved as efficient methods for stochastic RCPSP (SRCSP) where the uncertainty is modeled by probabilistic distributions, and the robustness is evaluated in terms of expected makespan. We cite the work of [11] in which metaheuristics were well investigated to SRCPSP. Recently, the work of [12] gives promising results for RCPSP under uncertainty.

In this context, we are encouraged to use the GRASP to the scenario-based robust RCPSP. Our tackled optimization problem aims to maximize the absolute robustness objective. We propose a GRASP algorithm enhanced with a forward-backward heuristic.

The next section focuses on the problem definition in deterministic and non deterministic version. Section 3 describes the main phases of the GRASP method. In section 4, we explicit the application of the latter method to the robust RCPSP. Computational results are given in section 5. Section 6 concludes the paper.

2. ROBUST PROJECT SCHEDUING PROBLEM

2.1. Deterministic RCPSP

A deterministic version of RCPSP consists in performing a set A of n activities on a set K of m resources. Every activity i has a fixed processing time denoted by p_i and requires r_{ik} units of resource type k which is characterized by a limited capacity R_k that must not be exceeded during the execution, and activities must not be interrupted. Two additive dummy activities 0 and n + 1 are used that represent to start and the end of the project, respectively. Dummy activities have null time duration and null resource requirement. The objective of the standard RCPSP is to construct a precedence and resources feasible schedule with the minimum makespan.

Precedence constraints perform that the start time of an activity *i* is permitted only when all its previous activities are finished. The Resource constraints satisfy that the use of every resource type, at every instant, does not exceed its capacity.

A schedule *S* referred to the baseline schedule which is given by the list of activity finish times (start times); let F_i (>=0) denotes the finish time of an activity *i*, then $S=(F_0, F_1, \ldots, F_n, F_{n+1})$ and the total project duration corresponds to the end project finish time F_{n+1}.

Therefore, the conceptual formulation of the RCPSP is given by the following formula:

$$\min F_{n+1} \text{ sc.} \tag{1}$$

$$F_h \le F_j - p_j; j = 1\Lambda\ n+1; h \in P_j \tag{2}$$

$$\sum_{i \in A(t)} ri_k \le R_k; k \in K; t \ge 0 \tag{3}$$

with $A(t)$ denotes the set of activities which are executing at time *t*, and P_j denotes the set of predecessors of the activity *i*.

An instance of the RCPSP can be represented by a graph G = (V, E) where the set of nodes V is defined by project activities and E contains arcs according to the precedence relations.

2.2. Min-Max robust RCPSP

The considered variability, for RCPSP under uncertainty, relies on activity durations. We use a scenarios-based approach to model the problem variability. Hence, we construct a set of scenarios, denotes by $\sum$, for optimization, let σ_i be a single scenario that corresponds to a problem realization. Each scenario is found by altering the initial activities durations with respect to a maximal activity delay.

A feasible solution *x* for the robust scheduling problem is represented by an activity list; let $f(x, \sigma_i)$ denotes the *makespan* of the generated schedule according to *x* on scenario σ_i. This value defines the local performance of the solution *x* according to σ_i. However, the global optimization process has to find the robust schedule with the global performance across the optimization set. Usually, the global performance is measured in terms of mean value, maximum deviation, etc.

The object of the present work is to optimize the min-max robustness objective of RCPSP which consists in minimizing the maximum makespan value over all scenarios. The optimization objective is given by the following formula.

$$\min \max_{\sigma_i \in \Sigma} (f(x, \sigma_i)) \tag{4}$$

Resource and precedence constraints for the robust RCPSP are the same in the deterministic case (Equations (2) and (3)).

2.3. Complexity

The robust scenario-based robust RCPSP with the min-max robustness objective is an NP-hard problem as it can be reduced to the standard NP-hard deterministic version for a number of scenarios equals to one [14].

3. GRASP METAHEURISTIC

The GRASP (Greedy Randomized Adaptive Search Procedures) is a multi-start metaheuristic which was developed for combinatorial optimization [15, 16]. It consists of an iterative process, in each of one, two phases are performed: a construction phase and a local search phase. The first one permits the construction of a feasible solution iteratively, one element at once iteration. However, the second phase performs the improvement of the recently constructed solution by a simple local search heuristic. The best across all generated solutions is then retained.

In the construction phase, a Candidate List (CL) is generated that contains the set of the candidate elements (edges) to be selected and added to the current partial solution. The CL is ordered with respect to a greedy function that measures the benefit of selecting each element. Moreover, the effective selection of one edge is done from an additive list: the Restricted Candidate List (RCL) that regroups the best elements from the CL with highest greedy values.

The GRASP procedure combines crucial characteristics of search methods. In the one hand, it is adaptive because the greedy function values are updated continuously depending on the current partial solution and the considered schedule construction strategy. In the other hand, it is a randomized-based method such that a selection of one element in the RC L is done randomly.

4. APPLICATION TO THE ROBUST RCPSP

We propose the application of the GRASP approach to the RCPSP with the optimization of the absolute robustness so called the min-max robustness objective.

The main steps of the proposed approach are depicted in the following figure. As a multi-start heuristic, the algorithm starts with generating gradually a new solution. This step integrates an intensification strategy. Current solution is improved, in the second step, by a Forward-Backward Improvement heuristic (FBI) and a Local Search heuristic (LS).

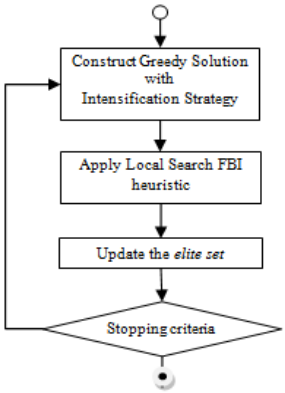


Figure 1. General steps of the enhanced GRASP approach

Throughout this iterative process an elite set (ES) is generated containing the best encountered solutions. The size of the ES is defined by a fixed parameter "*nbElite*". Elite solutions are updated iteratively.

4.1. Solution representation and *robust fitness*

A solution is represented by an activity list that satisfies precedence constraints. To evaluate the global solution performance, a *robust evaluation* is made. We apply decoding procedure to generate the schedule according to the activity list x and the scenario σ_i. Then, the *robust fitness* which measures the global solution performance is determined by the maximum makespan over all obtained values for $\sum$. The Serial Schedule Generation Scheme (Serial SGS) is used as a decoding procedure [2] to construct the schedule.

4.2. Construction phase with intensification strategy

At one iteration of the construction phase one activity is selected from an eligible set an added to the current partial solution. We generate the list CL of candidate activities having all their predecessors scheduled. For each activity in the CL, the corresponding greedy function value is equals to the priority rule value. We propose the application of different priority rules: the minimum Latest Finish Time (MLFT), the minimum of the activity free slack (MFLK), the inverse free slack priority rule, and the critical activity based selection. The object is to study the effect of the priority rule on robustness objective.

First activities of the CL are then copied in the *RCL*. The size of the latter list is denoted by *TRCL*. From the constructed RCL, an activity is then chosen randomly and added at the latest position in the partial activity list. A pool of elite solutions *ES* is constructed.

To ensure solutions with a high quality, we incorporate in the construction phase an intensification strategy based on the elite set.

In fact, when the *ES* attempts the fixed parameterized size, then, with probability *pES*, we select randomly one element to be considered at the current iteration of the construction phase. Then, the first activity, in the elite solution, that does not appear in the partial current solution is selected and inserted in.

The above described process is repeated until the construction of the totality of the solution is reached.

4.3. GRASP Improvement Phase

The proposed GRASP improvement phase combines a Local Search procedure (LS) with a FBI heuristic. The proposed Local Search starts from the recently constructed and improved solution x. Iteratively, a local move is applied to x to generate a neighbourhood set: N(x). The proposed move consists of the permutation of one activity of x with others nodes. The activity to be permuted is chosen at random. Obtained feasible solutions are saved to be compared with x. The best element over all neighbours and the current solution is retained. After a maximum number of iterations, the search method would stop with the best solution over all neighbourhoods.

The Forward-Backward Improvement (FBI) method is one of the basic heuristic for project scheduling. It was successfully hybridized with others methods ensuring efficiency and an acceptable computational time increment [17]. The Forward recursion is given by the serial SGS. However, the backward recursion is the SGS algorithm applied to the precedence-reverse graph starting from the end project activity where priority values are determined according to the lastly generated schedule.

5. EXPERIMENTS

5.1. Data Sets and Scenario Generation

The proposed approach was implemented in Java and ran on a portable personnel computer equipped with an Intel® Core™ i5-2450M CPU@ 2.50 GHz 773MHz, 2.70Go of RAM. Experiments were performed on two benchmark project instances: the Patterson data set [18], and the PSPLIB J30 data set [19]. The first data set contains 110 instances of various projects with 3 resource types and a number of activities that vary between 6 and 51. However, the second data set contains 480 project instances which are generated by the ProGen generator. These instances represent different projects with only 30 activities and 4 resource types.

We generated scenarios with limited size for both the optimization and the evaluation (simulation) set. The optimization set contains *nbScen* scenarios, equals to 10, used to compute the robustness objective. The evaluation set is used for simulation to estimate the expected makespan. The size of the evaluation set is denoted by *l*. A scenario is an initial problem realization where a set of activities are modified by altering their initial durations. In fact, for 10 percent of the total project activities, we add a time increment δ which is taken from a uniform distribution U(1, *maxDelay*). The latter parameter indicates the maximum activity delay which is fixed to 10.

In order to evaluate the performance of the proposed approach, we were interested by the following performance measures:

- The estimate *Expected makespan* which is calculated over the evaluation set $\left(\mathrm{E}(\mathrm{C}_{\max}) = \frac{1}{l}\sum_{i=1}^{l}(f(x,\sigma_i))\right)$;
- The *Standard deviation* of the makespan over the evaluation set;
- The *Relative Optimality gap* that measures the deviation between the estimate expected makespan and the lower bound *LB*, or the optimal makespan if exists, for the corresponding deterministic project $\left(\frac{E(\mathrm{C}_{\max})-LB}{LB}\right)$.

All results are averaged by the number of tested project instances.

5.2. Performance evaluation

5.2.1. Deterministic case

It is inevitable to study the algorithm behaviour on the deterministic case. Hence, the table 1 shows results of the GRASP implementation on the J30 data set. We performed the basic GRASP approach which based on a Local Search (LS) in the improvement phase (column 2). Then, the basic algorithm is improved with the FBI and tested for the same instances set. We vary the number of maximum iteration for both GRASP process (Line 2) and the local search (Line 3). Line 4 reported the average deviation from the well-known optimal solutions in percent, for both two GRASP implementations. The number of the obtained optimums is given in Line 5.

Table 1. Average deviations from optimal solutions J30 data set instances (the deterministic case).

	GRASP-LS	**GRASP-LS+BFI**			
Iterations for GRASP	100	100	300	3000	1000
Iterations for LS	100/2	10	10	10	1000/4
Optimality deviation	0.51	0.57	0.34	0.24	0.20
Optimums	396	390	419	428	434

Results for static RCPSP show the performance of the applied GRASP procedure compared with other methods in the literature [3], especially when combined with the forward-backward heuristic.

5.2.2. Results for Robust case

Under uncertainty, we have performed different runs of the proposed GRASP on Patterson data set. The basic algorithm denoted as (GRASP-LS(10)) is considered as the implementation of the GRASP approach with the LFT priority rule in the construction phase and 10 iterations of local search procedure.

We firstly, vary the maximum number of iterations with the local search incorporated in GRASP. Results are reported in table 2 with 1000 simulations as the size of the evaluation set. The evaluation procedure was ran for each project instance.

Table 2. Robustness evaluation on Patterson data set (1000 simulations).

	GRASP-LS(10)	**GRASP-LS(20)**	**GRASP-LS(100)**
Avg. Optimality gap	0.2366	0.2384	0.2249
Avg. Standard deviation	3.4423	3.4026	3.4121

Referring to the table 2, the local search procedure has an impact on the global performance. In fact, an increment of the number of iterations yields to better results for robustness. However, this

parameter must be controlled to ensure the non degradation of the later objective, which is the case with 100 iterations in the local search procedure. Table 3 contains numerical results of the proposed GRASP approach on J30 data set under uncertainty, simulated over 100 replications.

Table 3. Robustness evaluation on J30 data set (100 simulations).

		GRASP-LS(10)	**GRASP-LS(20) + BFI**
Nunber of iterations		500	250
Optimality gap	**avg.**	0.1349	0.1243
	max.	0.2755	0.2603
Standard deviation	**avg.**	4.2515	4.1428
	max.	6.3863	5.8086

In order to study the efficiency of the enhanced GRASP approach for the robust RCPSP, we evaluate the computational time on Patterson instances set for 1000 generated schedules, reported in Table 4. As described in [17], the FBI heuristics needs two passes of the SGS procedure to doubly justified the initial schedule. Thus, the number of generated schedules with the basic GRASP algorithm and the enhanced version with a FBI heuristic is equals to$(\text{nbIterMax} \times 10)$ and$(\text{nbIterMax} \times (10 + 3))$, respectively.

Table 4. Comparison between GRASP and GRASP-FBI on Patterson data set (1000 simulations).

		GRASP-LS(10)	**GRASP-LS(10) + BFI**	**GRASP-LS(10) + BFI**
Nunber of iteration		100	75	333
Optimality gap	**avg.**	0.2362	0.2304	0.2259
	max.	1.3352	1.31197	1.3398
Standard deviation	**avg.**	3.4106	3.4008	3.4089
	max.	5.5832	5.6207	5.5526
Time(s)		2.04	1.775	7.926

In table 4, column 2 and 3 show that for maximum 1000 generated schedules, the enhanced GRASP outperforms the basic GRASP in terms of robustness and computational time.

5.2.3. Priority rule

The idea of the present experiment is to study the effect of priority rules on robustness solution quality. As described in section 4, the construction phase implements a priority rule to order the Candidate List content, from which we select the *TRCL* best elements to the *RCL*.

We investigate in table 5 different priority rules based on critical path: Minimum Latest Finish Time (MLFT), the minimum Slack (MSLK), the inverse MSLK, the critical Activity based rule. The total activity slack is obtained by the difference between its latest and earliest start time. We also propose to study a priority rule which based on graph structure which is the GPRW (Greatest Rank Positional Weigth). We ran the GRASP algorithm with two different RCL size values (*TRCL*).

With limited size of the restricted list, the standard deviation of the estimated makespan is decreased as we reinforce best elements in the *RCL*. The inverse MSLK gives better results than the MSLK; this result can be interpreted as the inverse SLK favour activities having greatest slack values, consequently generated schedule will be more flexible to absorb activity delays.

Table 5. The Standard deviation variation on Patterson data set (1000 simulations).

	GRASP-LS(10)	
Priority rule	**TRCL=5**	**TRCL=3**
MLFT	3.4721	3.4456
MSLK	3.4702	3.4871
inverse MSLK	3.4640	3.4454
Critical Activity	3.4533	3.4741
GPRW	3.4606	3.4499

6. CONCLUSIONS

This paper has presented the application of the GRASP approach to the robust RCPSP. Based on scenarios, the object of the tackled optimization problem is to maximize the min-max robustnesss. The proposed GRASP approach incorporates priority rules in the greedy construction phase and two procedures in the improvement phase such as a local search and the forward-backward heuristic. Experiments have shown the simplicity of the GRASP implementation as a multi-start heuristic compared with other complex metaheuristic as evolutionary-based approach. In addition, the presented meta-heuristic was efficient to deal with uncertainty in acceptable computational time. Further works must be concentrated on the study of the GRASP construction phase to explore diverse solution on the search space, and the application of the algorithm to more large-sized project instances in robust scheduling.

REFERENCES

[1] Blazewicz, J., Lenstra, J., Rinnooy Kan, A., (1983), "Scheduling subject to resource-constraints: Classification and complexity", Discrete Applied Mathematics, Vol. 5, pp. 11–24.

[2] Kolisch, R., Hartmann, S. ,(1999), "Heuristic algorithms for the resource-constrained project scheduling problem: Classification and computational analysis", International Series in Operations Research and Management Science, Weglarz,J. (ed.), Vol. 14, pp. 147-178.

[3] Kolisch, R., Hartmann, S., (2006) ,"Experimental investigation of heuristics for resource-constrained project scheduling: An update", European Journal of Operational Research, Vol. 174, pp. 23-37

[4] Davenport, A.J., Beck, J.C., (2000), "A survey of techniques for scheduling with uncertainty", Available from http://tidel. mie. utoronto.ca/publications.php.

[5] Kouvelis, P. , Yu., G., (1997), "Robust Discrete Optimization and Its Applications". Klower Academic Publisher.

[6] Al-Fawzan, M., Haouari, (2004), "M.: A bi-objective model for robust resource-constrained project scheduling", International Journal of production economics, Vol. 18, pp. 1-13.

[7] Chtourou, H., Haouari, M., (2008), "A two-stage-priority-rule-based algorithm for robust resource-constrained project scheduling", Comput. Ind. Eng., Vol. 55, pp. 183-194.

[8] Artigues, C., Leus, R., Talla Nobibon, F., (2013), "Robust optimization for resource-constrained project scheduling with uncertain activity durations", Flexible Systems and Management Journal, Vol. 25(1-2), pp. 175-205.

[9] Sevaux, M. , Sorensen, K., (2004), "A genetic algorithm for robust schedules in a just-in-time environment", 4OR ,Quaterly journal of Operations Research Societies, Vol. 2(2), pp.129-147.

[10] Mogaadi, H., Fayech, B., (2015), "Scenario-Based Evolutionary Approach for Robust RCPSP", International Afro-European Conference for Industrial Advancement AECIA.

[11] Ballestin, F., R. Leus, (2009), "Resource-constrained project scheduling for timely project completion with stochastic activity durations", Production and Operations Management, Vol. 18, pp. 459-474.

[12] Creemers, S, (2015), "Minimizing the expected makespan of a project with stochastic activity durations under resource constraints", Journal of Scheduling, Vol. 18 (3), pp. 263–273.

[13] S. Horroelen, Leux, R.., (2005), "Project scheduling under uncertainty: Survey and research potentials", European Journal Of Operational Research, Vol. 165(2), 289–306.

[14] Aissi, H., Bazgan, C., Vanderpooten, D. (2009), "Min–max and min–max regret versions of combinatorial optimization problems: A survey", European journal of operational research, Vol. 197(2), pp. 427-438.

[15] Feo, T.A., Resende, M.G.C., Smith, S., (1994), "A Greedy Randomized Adaptive Search Procedure for Maximum Independent Set", Operations Research, Vol. 42, pp. 860-878.

[16] Festa, P., Resende, M. G. (2009), "Effective application of GRASP", Wiley encyclopedia of operations research and management science.

[17] Valls, V., Ballestin, F., Quintanilla, S., (2005), "Justification and rcpsp: A technique that pays", European Journal of Operational Research, Vol. 165, pp. 375-386.

[18] Patterson, J. H.,(1984) ,"A comparison of exact approaches for solving the multiple constrained resource, Project Scheduling Problem", Management Science, Vol. 30, p854-867.

[19] Kolisch, R., A. Sprecher, (1996), "PSPLIB - A project scheduling problem library", European Journal of Operational Research, Vol. 96, pp. 205-216.

AUTHORS

Hayet Mogaadi received a diploma of Engineer in Computer Science from the National School of Computer Sciences (Tunisia) in 2003, and a Master degree in Automatic and Signal Processing from the National Engineering School of Tunis (Tunisia) in 2005. She is a Ph.D. student in Electrical engineering at the National Engineering School of Tunis. Her interest's area is project scheduling.

Besma Fayech Chaar received the diploma of Engineer in Industrial Engineering from the National Engineering School of Tunis (Tunisia) in 1999, the D.E.A degree and the Ph.D degree in Automatics and Industrial Computing from the University of Lille (France), in 2000, 2003, respectively. Currently, she is a teacher assistant at the University of Tunis (Tunisia). Her research interests include artificial intelligence, decision-making systems, scheduling, and transportation systems.

INVESTIGATING THE INFLUENCE OF SERVICE TRAINING, REWARD SYSTEM AND EMPOWERMENT ON JOB SATISFACTION AND ORGANIZATIONAL COMMITMENT OF THE EMPLOYEES OF ILAM'S TELECOMMUNICATIONS COMPANY

Mohammad Taban[1], Seidmehdi Veiseh[2] and Yasan allah Poorashraf[3]

[1]Assistant Professor of Management Department, University of Ilam
taab1351@ut.ac.ir
[2]Assistant Professor of Mathematic Department, University of Ilam
Amir7912000@yahoo.com
[3]Associate Professor of Management Department, University of Ilam
yasan_ashraf@yahoo.com

ABSTRACT

The current study focuses on the investigation of the relationship between variables training, reward, empowerment and job satisfaction of the employees and the influence of job satisfaction on the organizational commitment. In so doing, the relationship between super pattern variables was considered on the basis of some theoretical principles. This is a correlative study in which the path analysis model is used, because it is possible to investigate direct and indirect paths through such model. Population included all employees of Ilam's telecommunications company among whom 190 ones were selected as the sample by use of randomized sampling model. Data was gathered using five standard questionnaires. The reliability and admissibility of the questionnaires was evaluated by use of Coronbach'sα and explorative factor analysis. In order to test the available pattern, the path analysis technique was used. Results show that there is a meaningful relationship between the above mentioned variables.

KEYWORDS

Training, reward, empowerment, job satisfaction, organizational commitment, Telecommunications Company

1. INTRODUCTION

During the past decades, an organization was mainly evaluated on the basis of factors such as financial resources and equipment. But today, the main comparative capitals of the organizations are the skilled and educated persons who are thoroughly proficient in the technological

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 147–156, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60213

knowledge. In comparison to 1982, the big organizations are the owners of %15 of tangible and %85 of intangible assets of the world (Azarhoosh, 2005). One of the main mental disturbances of the managers is productivity and the factors affect on. Althin (2005) believes that productivity of each organization depends mainly on its human resources. The Japan's productivity association has declared that the productivity of each organization is affected by the factors such as empowerment of the employees, training, participative management, justice and just distribution (Steiner, 1997). It is hypothesized that there is a direct relationship between effective use of human resource and increase of organizational commitment (Khaki, 2010). Under the influence of not taking into consideration the human resource management of the organization, the employees' loyalty and commitment is decreased. Telecommunications Company is also affected by the human resource capital. In this regard, it is important to obviate the employees' problems and motivate them in order to increase the quality and performance of the company. In this study, the organizational commitment of the employees of Telecommunications Company is regarded as the dependent variable affected by factors such as training, reward, and empowerment and job satisfaction.

2. STATEMENT OF ISSUE

Today, the organizations should be managed in a comparative environment. In such conditions, the mangers have no suitable opportunity to managing the employees who are charged with the most of duties. The employees are able to perform delivered duties when they enjoy suitable skills, knowledge and abilities. By empowerment of the employees, we mean a set of systems, approaches and measures used for increasing productivity of the organization and human resources (Carter, 2001). The word "empowerment" is referred to the person's knowledge from him/her self. Empowerment is related to the factors such as organizational environment, the relationship between employees, manager and coworkers (Moye, & Henkin, 2006).. It is defined as one of the main management procedures for making decisions (Melham, 2004). Empowerment of the employees refers to when the employees can develop their abilities and knowledge in order to enhance personal and organizational goals. In reality, empowerment is a process through which the internal opinions of the persons are changed (Vtn, &Cameron, 1991). Reviewing the literature, Karakoc (2009) argues that the employees' empowerment is influenced by some factors including inclination for developing , having the art of criticism , acceptance of change, high self – confidence, dynamic structure, evaluation of performance, feedback, reward, support, training, communications, motivation, encouragement, participative culture, flexibility, freedom of speech, information sharing and management confidence. In order to perform empowerment, the managers should equip the employees with needed information and training. The managers should become sure that the employees receive continual training (Senate et al, 2007). Training is aimed to solve the organization's problem, create necessary changes, increase the knowledge of new employed staff, share information and develop skills (Berge, 2008). Moreover, it is possible to solve some problems for the employees by training (Hosseinzadeh&Barziagar, 2004).

On the other hand, Rand Lip states that the mangers have a tendency toward the empowerment of the employees. Moreover, the employees suppose them valuable and have more participation in working. Empowerment entails positive results including workforce replacement, productivity improvement and increase of organizational commitment and motivation (Rasouly, 2005). Furthermore, empowerment leads to increase the performance of various units of the organization, the mutual respect of the employees and the emphasis on how the problems are

solved (Klug, 1998). Job satisfaction is one of the main consequences of empowerment. It has been said (Tomas&lighthouse) that high self-confidence can increase the job satisfaction.

Moreover, freedom of action can also lead to the increase of job satisfaction. Telecommunications Company is one of the main companies of each country. Therefore, investigation of factors such as job satisfaction and organizational commitment of the employees is of great importance. Moreover, it is important to recognize the variables having influence on the employees' job satisfaction and commitment, because productivity of Telecommunications Company is affected by such variables.

3. RESEARCH GOALS

This study aims to investigate the relation between training, reward, empowerment and job satisfaction of the employees and the influence of job satisfaction on the organizational commitment. In this regard, the followings should be mentioned:

1. The relationship between training and job satisfaction of the employees of Ilam's telecommunication company

2. The relationship between reward system and job satisfaction of the employees of Ilam's telecommunication company.

3. The relationship between empowerment and job satisfaction of the employees of Ilam's telecommunication company.

4. The relationship between job satisfaction and organizational commitment of the employees of Ilam's telecommunication company.

4. RESEARCH HYPOTHESIS

Hypothesis 1: there is a positive and meaningful relationship between training and job satisfaction of the employees of Ilam'stele communications company.

Hypothesis 2: There is a positive and meaningful relationship between reward and job satisfaction of the employees of Ilam's telecommunication company.

Hypothesis 3: there is a positive and meaningful relationship between empowerment and job satisfaction of the employees of Ilam's telecommunications company.

Hypothesis 4: there is a positive and meaningful relationship between job satisfaction and organizational commitment of the employees of Ilam's telecommunications company.

5. METHOD

Since the relationship between the variables is analyzed on the basis of research goal, this is a correlational – descriptive research. This study is aimed to recognize the cause and effect patterns

of the variables. This approach allows the researchers to analyze the possible relationship between the variables of the study. Data was analyzed by use of structural equations modeling.

5.1 Population

Populations included all employees of Ilam's telecommunication company in 2013-2014. They were selected by use of simple randomized sampling method.

5.2 Sample

Determination of sample of the research is one of the main problems to which the authors are always faced. It has been decided to consider 8 persons for each parameter. Therefore, the sample of the research includes 200 employees.

6. FINDINGS

Table 1: correlation matrix of the research variables

	1			
Reward	/23	1		
Empowerment	/22	/37	1	
Job satisfaction	/28	/38	/71	1
Organizational commitment	/27	/33	/52	/61

Correlation matrix is used for examining the linear relationship between the variables. The figures varied from -1 to +1 and the high degree of the correlation reveals the ratio of the relationship between the variables. Correlation matrix related to the variables is seen in the above table.

As shown in the above table, there is a meaningful correlation between the variables. This is, therefore, to say that there is a linear relationship between the variables. Moreover, the available coefficients are between %20 and %90; it means that there is no meaningful relationship between the variables.

Table 2: The HomographSmirnoff's test used for investigating normal condition of the variables

		Compensation of services	Reward	Empowerment	Organizational commitment	Job satisfaction
Normal parameters	**Average**	3.23	3.23	3.22	3.19	3.31
	Standard deviation	0.344	0.507	0.502	0.382	0.582
Clomograph – Simronoph's factor		2.91	1.40	2.77	1.73	2.64
Meaningful level		0.098	0.089	0.077	0.051	0.118

Table above shows that data distribution is normal (p>/5) and it is possible to ignore the linear correlation between the variables. Regarding the hypothesis of path analysis, the following figure is presented for hypothesis test.

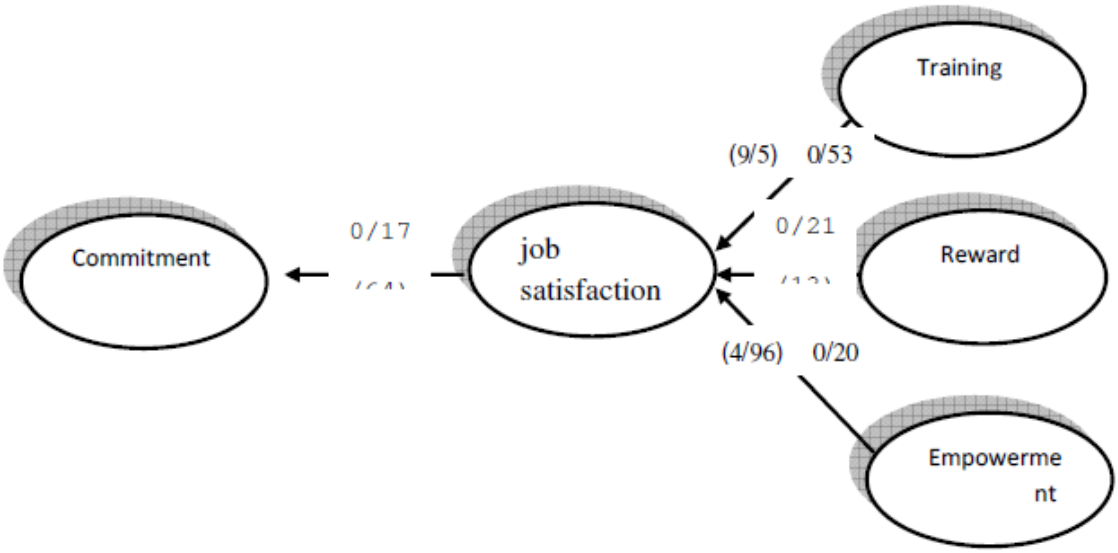


Figure 4-1: hypothesized model after fitting

As seen in the above figure, all paths are statistically meaningful. These paths include: (a) the path of influence of training on job satisfaction (β=/53, t= 9/5); (b) the path of influence of reward on job satisfaction(β=/21, t= 5/13);(c) the path of influence of empowerment on job satisfaction (β=/20, t= 4/96) and(d) the path of the influence of job satisfaction on organizational commitment (β=/17, t= 3/64). Moreover, all parameters show the positive and meaningful effect of the variables.

7. FINAL MODEL OF THE RESEARCH

a) Direct, indirect and total affect of exogenous variable on the hidden endogenous variables

In final model of the research, all direct paths of influence of hidden exogenous variables on the hidden exogenous variables on the hidden endogenous ones are investigated. In this regard, it is necessary to examine direct influence of hidden exogenous variables (training, reward and empowerment) on the hidden endogenous variables (job satisfaction and organizational commitment). The parameters related to the direct influence of exogenous variables are seen in the table below.

Table 3: Parameters related to the direct / indirect influences of hidden exogenous variable on the hidden endogenous ones

Hidden exogenous	**Effects**	**Hidden endogenous / parameters**	**Job satisfaction**	**Organizational commitment**
Training	Direct effectiveness	Parameter b estimation	0.29	-

		Standard parameter B	0.53	-
		Standard deviation	0.50	-
		T	9.5	-
	Indirect effectiveness	Parameter b estimation	-	0.73
		Standard parameter B	-	0.43
		Standard deviation	-	0.06
		T – value	-	11.66
	Total effect	Parameter estimation	0.29	0.73
		Standard parameter B	0.53	0.43
		Standard deviation	0.05	0.06
		T – value	9.5	11.66
Reward	Direct effects	Parameter estimation	0.17	-
		Standard parameter B	0.21	-
		Standard deviation	0.08	-
		T –value	5.13	-
	Indirect effects	Parameter b estimation	-	0.38
		Standard parameter B	-	0.42
		Standard deviation	-	0.06
		T – value	-	7.66
	Total effects	Parameter b estimation	0.17	0.38
		Standard parameter B	0.21	0.42
		Standard deviation	0.08	0.06
		T – value	5.13	7.66
Empowerment	Direct effects	Parameter b estimation	0.27	-
		Standard parameter B	0.20	-
		Standard deviation	0.08	-
		T – value	4.96	-
	Indirect effects	Parameter b estimation	-	0.44
		Standard parameter B	-	0.41
		Standard	-	0.05

		deviation		
		T – value	-	8.90
	Total effects	Parameter b estimation	0.27	0.44
		Standard parameter B	0.20	0.41
		Standard deviation	0.08	0.05
		T – value	4.96	8.90

As seen above, the t – value related to the all direct influences of hidden exogenous variables on endogenous variables is 2 and it is statistically meaningful. The biggest figure is related to the direct influence of training on job satisfaction (β=/53) and the weakest is related to the influence of empowerment on job satisfaction (β=/20).

Finally, total affect include the sum of direct and indirect affects of the variables. Moreover, all t – values are equal to %5. The most powerful influences are the influence of training on job satisfaction (β=/53). And the influence of training on organizational commitment (β=/42).

b) Direct influence of hidden endogenous variables on other endogenous variables

Table 4: parameters related to the direct influences of hidden endogenous variable (Y)

Hidden endogenous	**Hidden endogenous**	**Parameter B estimation**	**Standard parameter B**	**Standard deviation**	**T – value**
Job satisfaction	Organizational commitment	0.13	0.17	0.04	3.64
P<%5					

Findings show that the t – values relative to the relation between job satisfaction and organizational commitment are meaningful. Therefore, the endogenous variable job satisfaction influences directly on the variable organizational commitment. From the above table, β=/17 and its error degree is %4.

c) Total influence of hidden endogenous variables on each other

Parameters related to the total influence of hidden endogenous variable job satisfaction on the other variable organizational commitment are shown in the table 5

Hidden endogenous	**Hidden endogenous**	**Parameter B estimation**	**Standard parameter B**	**Standard deviation**	**T – value**
Job satisfaction	Organizational commitment	0.13	0.17	0.04	3.64
P<%5					

Finding show that when p<%5, the t – value related to the influence of job satisfaction on organizational commitment is meaningful.

In reality, the total influence of the endogenous variable job satisfaction on organizational commitment is statistically meaningful, because β=/17 and error degree is %4.

d) The variance of hidden endogenous variables

After investigating the direct, indirect and total influences of the model, it is necessary to determine the variance of organizational commitment on the basis of endogenous and exogenous variables..

Table 6: the variance determined on the basis of hidden endogenous variables

Anticipative variables	**Anticipated variables**	**R^2**
Training	Job satisfaction	0.28
Training and job satisfaction	Organizational commitment	0.46
Reward	Job satisfaction	0.75
Reward and job satisfaction	Organizational commitment	0.59
Empowerment	Satisfaction	0.45
Empowerment and satisfaction	Organizational commitment	0.25

Due to the findings, the variable organizational commitment is mainly affected by the variables reward and job satisfaction (R^2=/59). This is to say that based on the final model of the research, the variance of educational development is determined by %25.

8. DISCUSSION AND CONCLUSION

Hypothesis 1: the variable training affects directly on job satisfaction of the employees of Ilam'stelecommunications company. Results showed that job satisfaction of the employees is directly affected by the training.

Since the improvement of job knowledge is one of the main feedbacks of training periods, it seems that job satisfaction of the employees of Telecommunications Company is mainly affected by training. The employees of an organization can perform their duties better when they receive training programs.

Hypothesis 2: reward system affects directly on job satisfaction of the employees.

From the results, it becomes clear that job satisfaction is directly and meaningfully affected by reward system. It is hypothesized that the employees of an organization put emphasize on factors such as salary, advancement, work identity, organizational procedures and work conditions. In this regard, the managers should prepare the fields of satisfaction of the employees, because there is a meaningful correlation between job satisfaction and the employees' effectiveness and mental health. In other words, the employees work better when they are supported by the organization. On the other hand, the organization's support has influence on the employees' self confidence.

Hypothesis 3: empowerment has direct and meaningful influence on the employees' job satisfaction.

Since β=/20 and t – value= 4/96, job satisfaction is meaningfully affected by empowerment.

As mentioned before, empowerment is one of the modern approaches by which the employees are motivated to perform their duties better. It is argued that empowerment – based management increases the human resources' effectiveness. The managers should make clear the employees' responsibilities and objectives by making participative decisions. From organizational perspective, the human resources sector should make developmental programs available to the employees. In other words, the managers should have control over the employees' performance and cause them to experience the feeling of joy and respect. Moreover, organizational productivity and commitment is increased by moral values of the staff, including confidence and validity.

Hypothesis 4: job satisfaction affects directly and meaningfully on the employees' organizational commitment.

Evidences show that organizational commitment of the employees is directly affected by job satisfaction. Moreover, organizational dynamism depends on factors such as the presence of skilled and creative persons. The employees should be loyal to the organization. The loyalty of the employees causes more organizational development.

It is important to investigate the relationship between job satisfaction and organizational commitment as the two variables having influence on the development of organizational goals and the employees learning. The employees can help the organization to fulfill its objectives. Through being satisfied, the employees can help the organization to fulfill its goals. The employees of Telecommunications Company try to preserve their job opportunities because transfer of learning is impossible. Such behavior has negative influence on the relations between the employees and the managers. Therefore, the managers should allow the employees to participate in decision makings in order to stabilize their positions. They should make clear the employees' role in order to increase their affiliation and effectiveness.

REFERENCES

[1] Althin, L. (2005). Efficiency And Productivity of Employment Offices Employment: Evidence From Sweden,. International Journal of Manpower, 26(2), 196-206.

[2] Azarhosh, F (2005), Trends and new ideas in management, Proceedings of the Second Conference on Human Resource Development. Industry Development and Renovation Organization of Iran.

[3] Carter, M. (2001). Strategic planning in nonprofit or for profit organizations. from www.Strategic planning .com

[4] Karakoc, N. (2009). Employee Empowerment and Differentiation in Companies: A Literature Review and Research Agenda. Enterprise Risk Management, 1(2)

[5] Khaki, G. (2010). Approach to the research dissertation. Tehran, publishing reflection.

[6] Klagge,J(1998) The empowerment squeeze-views from the middle management position. Journal of Management Development, Vol 17. No 8. P 548-558.

[7] Hossein-Zadeh, D., Barzegar N. (2004), Learning processes in organizations. Islamic Azad University. Save

[8] Moye, M. Henkin, A(2006) Exploring association between employee empowerment & interpersonal trust in manager. Journal of Management Development. Vol 25. No 2. P 101-117.

[9] Melhem,Y(2004) The antecedents of customer-contact employees empowerment. Employee Relations,Vol 26. No 1.p 72-93

[10] Rasouli, R. (2005) ,examined the relationship between employee empowerment, job stress, job satisfaction and organizational commitment. Quarterly Message. No. 16-15. Pp. 194-165

[11] Seadat talab, A, Yasin, A. (2011). The relationship between quality of work life of faculty motivation and martyr Beheshti University in Tehran. National Conference on Higher Education Isfahan.

[12] Stainer, A. (1997). Logistic-a Productivity and Performance Perspective. Suuly Chan Management an International Journal, 2(4), 53-62.

[13] Thomas, K.W. and Velthouse, B.A. (1990). Cognitive elements of empowerment: an interpretive model of intrinsic task motivation'' Academy of Management Review, Vol. 15 No. 4, pp. 666-81

[14] Vtn, D. Cameron, K, (1999, empowerment and delegation, Translation: . Avry Yzdanyv Badroddin ,Management Education Research Institute.

ASSESSING THE SKILLS AND COMPETENCIES AMONG PRINCIPALS IN TELECOMMUNICATION COMPANY LOCATED IN ILAM

Seidmehdi Veiseh[1], Yasan allah Poorashraf[2], Mohammad Taban[3]

[1]Assistant Professor of Management Department, University of Ilam,Iran
Amir7912000@yahoo.com
[2]Associate Professor of Management Department, University of Ilam
yasan_ashraf@yahoo.com
[3]Assistant Professor of Management Department, University of Ilam
taab1351@ut.ac.ir

ABSTRACT

The main objective of the present study was to identify the required skills and competences for the managers of Telecom Company of the city of Ilam from the employee's perspective. The research method was descriptive-survey, and the participants were the staff of Ilam Telecom Company. 190 participants were selected randomly. The results of t-test and one the way ANOVA at the level of 0.05, and 95% confidence level showed that there is no significant difference between academic discipline and cognitive skills. There was a significant difference between gender and the three skills. There is no significant difference between experience and conceptual skills. Management competence was obtained based on the t-test and analysis of variance of these results. There was significant difference between gender and level of the managerial competence. There was a significant difference between the field of the study and the level of the managerial competence. And finally, between level of education and level of the conceptual competence, there was no significant difference.

KEYWORDS

Skills, managerial skills, management competence, technical skills, conceptual skills, human skills

1. INTRODUCTION

Today, human resources are an organization's most important asset, so if the conditions for the exercise of managerial skills changes, the proper coordination should be made by the managers (Hammer & champy, 2005). To achieve this, managers must take appropriate actions and coordination and put their training to the test. Usually managers are responsible for planning, organizing, leading, and controlling the various stages of an organization (shenhar & Reiner, 2001). Managers are like people who are always on the hunt for opportunities to develop and enhance their knowledge, and to establish a strong relationship using their various skills. Drucker

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 157–166, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60214

(2003) has argued that the challenges that occur in an organization are reflected in the implementation of effective and expansive changes, in this respect, the cooperation of managers paves the way for the position of an organization. This was conducted by the grounding of the character traits and skills of successful managers, and the methods of managing (Katzenbach & smith, 2003). Another issue in this research is the concept of competence. Competencies are the axis of the heart of assessment and other elements are formed by them. Applying the term, competency, is a new approach that emerged in the 1970s, and developed and expanded in later years. The meaning of competence is the knowledge, abilities, skills, attitudes, and motivations for doing a successful work.

2. STATEMENT OF PROBLEM

As mentioned above, management level, type and position of the organization, and demands placed on it are the major factors in skills, abilities, and competences required for managing. Consequently, with the knowledge of the effect of each of these three elements in needed combined skills, we can achieve educational needs and moving system of the manager in the management hierarchy. In recent centuries, the strong tendency for staff training has resulted in significance development of training in organizations and it has changed the opinion of the community about the missions and responsibilities of each organization completely. Today, the organizations have a more important role and the role of the managers is also more important because of the special role that each organization has in the developing the culture, society, and economy, it is more important than other occupations (Moayeri, 1998). The managers are under constant analysis because of the professional characteristics, assisting the difficult people (who are very resistant), non- voluntary clients and resolving tensions between employees. In most organizations, many problems are caused by the methods of management . Meanwhile, the good management of the organization depends on delegation and on the other hand, depends on the method used, managerial skills, and competence of managers. In categorizing the problems and the methods of using of skills and management competence by the managers are part of the operating issues in the management of the organizations. The majority of the managers of telecom organizations have relatively low technical skills and in some cases low human conceptual skills. Consequently, this research tries to study the skills and competencies required for managers of Ilam Telecom Company from the employees' point of view. So, this study seeks answer whether managers of Telecom Company have the required skills and competencies? And whether their management skills and competencies are different according to key demographic variables?

3. HYPOTHESIS

3.1 General hypothesis:

- Telecom Leaders possess desirable key skills and professional competence.

3.2 Specific hypotheses:

- Telecom managers possess desirable technical skills.
- Telecom managers possess desirable human skills
- Telecom managers possess desirable conceptual skills.

- Managers possess desirable key competence (academic, managerial, conceptual, and moral competence).

4. CERTAIN CONCEPTS

Skills: the fosterable ability of a person that is reflected in performing and playing role and its main measure is performance and action in various conditions. (Alagheband, 2005)

Competence: describes a set of the behaviors that reflects a unique combination of knowledge, skills, abilities, and motivations and is associated with the performance in an organizational role (Viitala, 2005).

Managerial skills: management skills, a set of behaviors that leads to effective performance in a job.

Technical skills: the knowledge and ability to perform specific tasks that requires mastery of techniques and special tools, and practical competence in behavior and activities (Alagheband, 2005).

5. METHODOLOGY

The present study is a practical study with regard to its aim and a descriptive study with regard to the type of study. Evaluative studies are a form of scientific research that evaluate the effectiveness or practicality of changed programs (Saei, 2008). In this study, like a survey, the researcher uses questionnaire or interview to explore ideas, thoughts, perceptions and preferences of its intended people. Necessary data to test the research hypotheses were collected by a questionnaires given to a sample and was analyzed after filling the questionnaires.

• The participants

The participants of the present study were the entire staff of Ilam Telecom Company. It should be noted that in this study we only considered the employees that by the year 1392 had at least 3 years working experience. Thus, the study population consisted of 377 employees.

• Sampling method

In this study the simple random sampling was used. Random sampling is a method to choose a part of a society in a way that all the possible samples have an equal chance of being chosen (Khaki, 2011).

• The sample size

There are many methods and formulas to determine sample size. In this study, the sample size was calculated using Cochran formula. The number of cases per 377 was estimated 190 people.

6. THE ANALYSIS OF RESEARCH HYPOTHESIS

Hypothesis 1: The attitude of both males and females is different regarding management skills.

To assess this hypothesis, the t-test was used for comparison of two groups of males and females. The findings are as follows.

Table 1. t-test to compare means between two groups

T-test for equality of variances							Leven test for equality of variances			
With 95% difference confidence between repeated t-test		Standard deviation	Deviation from the mean	Significance level	Degrees of freedom	t	Sig.	F		
most	least									
0.000	-0.108	0.028	-0.054	0.042	188	2.94-	0.039	0.711	Assuming equal variances	Communicative skills
0.000	-0.108	.0.028	-0.054	0.041	187.7	2.95-			Assuming inequality - Variance	
0.020	-0.076	0.025	-0.028	0.025	188	3.14-	0.029	1.101	Assuming equal variances	Technical skills
0.020	-0.076	0.025	0.028	0.023	187.7	3.15-			Assuming inequality - Variance	
0.085	0.011	0.019	0.048	0.017	188	2.57	0.024	5.089	Assuming equal variances	Human skills
0.085	0.011	0.019	0.048	0.014	187.7	2.57			Assuming inequality	

the formulation of hypothesis ✓

$H0 : \mu 1 = \mu 2$ ✓

$H1 : \mu 1 \neq \mu 2$ ✓

calculation of t: = 2.57 -3.14, t3 = t1 = -2.94, t2
Determining the degrees of freedom: df = 188
T Extractor table with degrees of freedom 188, t = 1.96
T table> the calculated t and Comparison and Conclusion: for the three skills

Based on the Leven test, the significance level was less than 0.05 for all three types of skills (Sig1 = 0.03, Sig2 = 0.029, Sig3 = 0.024), therefore we used the results which did not assume equal variances. Since the Leven test significance level is less than 0.05, we applied the second test, t-test for two independent groups.

As presented in the above table, the calculated t value of 2.94 was obtained for conceptual skills (of course, taking the absolute value of t), on the other hand, the value of t table with degrees of

freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 there is a significant difference between the attitudes of male and female employees, regarding the communicating skills of managers. On the other hand, Sig is lower than 0.05, so we can confirm the hypothesis with 95% confidence.

T value of 3.14 was calculated for technical skills (of course, taking the absolute value of t), on the other hand, the value of t table with degrees of freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 there is significant difference between the attitudes of male and female employees, regarding the technical skills of managers. On the other hand, Sig is lower than 0.05, so we can confirm the hypothesis with 95% confidence.

.

T value of 2.57 was calculated for human skills (of course, taking the absolute value of t), on the other hand, the value of t table with degrees of freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 there is significant difference between the attitudes of male and female employees, regarding the technical skills of managers. On the other hand, Sig is lower than 0.05, so we can confirm the hypothesis with 95% confidence.

The second hypothesis: the views of people according to their field of study are different about the managerial skills of the managers.

Table 2. the independent t-test to compare the means of two groups of managerial and non-managerial

T-test for equality of variances							Leven test for equality of variances			
With 95% difference confidence between repeated t-test		Standard deviation	Deviation from the mean	Significance level	Degrees of freedom	t	Sig.	F		
most	least									
0.122	-0.029	0.038	0.046	0.130	188	1.202	0.023	0.327	Assuming equal variances	Communicative skills
0.199	-0.027	0.037	0.046	0.120	187.3	1.246			Assuming inequality - Variance	
0.152	0.018	0.034	0.085	0.012	188	2.50	0.032	3.699	Assuming equal variances	Technical skills
0.157	0.013	0.034	0.085	0.021	187.3	2.344			Assuming inequality - Variance	
0.021	-0.081	0.026	-0.035	0.243	188	-1.15	0.040	0.028	Assuming equal variances	Human skills
0.024	-0.083	0.027	-0.035	0.243	187.3	-1.10			Assuming inequality	

the formulation of hypothesis

$$\begin{cases} H0: \mu 1 = \mu 2 \\ H1: \mu 1 \neq \mu 2 \end{cases}$$

Calculating the value of t :=-1.15 2.50 , t3 = t1= 1.202 , t2
Determining the degree of freedom: df=188
T Extractor table with degrees of freedom 188, t = 1.96

Comparison and conclusion: for the technical skill the calculated t is > t table and for the other two skills the relation is reversed.

Based on the Leven test, the significance level was less than 0.05 for all three types of skills (Sig_1=0.023, Sig_2=0.032, Sig_3=0.040), therefore we used the results which did not assume equal variances. Since the Leven test significance level is less than 0.05, we applied the second test, t-test for two independent groups.

As presented in the above table, the calculated t value of 202.1 was obtained for conceptual skills (of course, taking the absolute value of t), on the other hand, the value of t table with degrees of freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 there is no significant difference between the attitudes of employees, regarding the communicating skills of managers. On the other hand, Sig is higher than 0.05, so we can reject the hypothesis with 95% confidence.

T value of 50.2 was calculated for technical skills, on the other hand, the value of t table with degrees of freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 that regarding the employees' point of view there is significant difference between the technical skills of managers and their field of study.

T value of - 15.1 was calculated for human skills, on the other hand, the value of t table with degrees of freedom of 188 is equal to 1.96, therefore because the calculated t is greater than t table, we can claim that in the confidence level of 0.05 regarding the employees' point of view there is significant difference between the human skills of managers and their field of study. On the other hand, Sig is higher than 0.05, so we can reject the hypothesis with 95% level of confidence.

The third hypothesis: The view of staff is different about the level of management skills of managers according to their level of education

.

Table 3. One-way analysis of variance

Sig.	**F value**	**Mean square**	**Degrees of freedom**	**Sum of squares**	**Source of change**	**Variables of the research**
0.045	2.69	0.355	3	1.064	Between-group differences	Communicative skills
		0.132	186	90.73	Intra-group differences	
		-	189	91.73	total	
0.029	3.032	0.312	3	0.937	Between-group differences	Technical skills
		0.103	186	70.86	Intra-group differences	

		-	189	71.80	total	
0.135	1.86	0.113	3	0.339	Between-group differences	Human skills
		0.061	186	41.84	Intra-group differences	
		-	189	42.08	total	

the formulation of hypothesis

$$\begin{cases} H0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H1 : \mu 1 \neq \mu 2 \neq \mu 3 \neq \mu 4 \end{cases}$$

Calculation of f value := 1.86 3.032 , f3 = f1= 2.69, f2
Determining the degree of freedom: the row degree of freedom 3 and column 186
extraction from Table f: F (df: 3, 186 & p = 0.05) = 2.62
comparison: f3 < fj , f2 > fj, f1 > fj

Conclusion: As presented in the above table, because the calculated f value for technical and conceptual skills was higher than f table, so we can claim that with the confidence level of %95 that there is a significant difference between the means of educational groups regarding the technical and conceptual skills. On the other hand, Sig for the two skills was lower than 0.05, (Sig_1= 0.045, Sig_2= 0.029), so again, with the level of 95% confidence we can claim that there is significant difference between the educational groups regarding the assessing the technical and conceptual skills of managers.

Regarding the human skills, because the calculated f is lower than the f table, we can claim with the level of confidence of %95 that there is no significant difference between means of the educational groups regarding the assessing the human skills of managers. On the other hand, as the table shows Sig. is higher than 0.05. So, again with the level of confidence of %95 we can claim that there is no significant difference between the level of education of the employees regarding the assessing the human skills of the managers.

The structure graph of differences which is about the comparison of means suggests that where differences may occur. As the structure graph of differences indicates when the degree is higher than Bachelor, the difference between the various types of skill becomes greater.

The fourth hypothesis: Based on experience, perceptions of the staff about management skills differs.

Table 6-4. One-way analysis of variance

Sig.	**F value**	**Mean square**	**Degrees of freedom**	**Sum of squares**	**Source of change**	**Variables of the research**
0.102	1.93	0.25	4	1.02	Between-group differences	Communicative skills
		0.13	185	90.77	Intra-group differences	

		-	189	91.79	total	
0.001	4.80	0.48	4	1.95	Between-group differences	Technical skills
		0.10	185	69.85	Intra-group differences	
		-	189	71.80	total	
0.004	3.84	0.23	4	0.92	Between-group differences	Human skills
		0.06	185	41.16	Intra-group differences	
		-	189	42.08	total	

the formulation of hypothesis

H0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H1 : $\mu 1 \neq \mu 2 \neq \mu 3 \neq \mu 4 \neq \mu_6$

Calculation of f value := 3.840 4.804 , f3 = f1= 1.939 , f2

Determining the degree of freedom: the row degree of freedom 4 and column 185

Extraction from Table f: $F_{(df: 4, 185 \& p=0.05)} = 2.39$

comparison: f3 > fj , f2 > fj , f1 < fj

Extraction of conclusion: As presented in the above table, because the calculated f value for technical and conceptual skills was higher than f table, so we can claim that with the confidence level of %95 that there is a significant difference between the means of years of experience regarding the technical and conceptual skills of the mangers. On the other hand, Sig for the two skills was lower than 0.05, (Sig3= 0.004, Sig2= 0.001), so again, with the level of 95% confidence we can claim that there is a significant difference between the years of experience regarding the assessing the technical and conceptual skills of managers.

Regarding the conceptual skills, because the calculated f is lower than f table, we can claim that with the confidence level of %95 there is a significant difference between the means of years of experience regarding the conceptual skills of the mangers. On the other hand, as presented in the table Sig for the two skills was higher than 0.05. So again, with the level of 95% confidence we can claim that there is no significant difference between the years of experience regarding the assessing the technical and conceptual skills of managers.

The structure graph of differences which is about the comparison of means suggests that where differences may occur. As the structure graph of differences indicates the higher the years of experience, the better the staffs' assessment is about technical and human skills of the managers. Difference between the various types of skill becomes greater. The structure graph of differences is as follows. And conceptual skills has decreased-increased changes

6. SUMMARY AND CONCLUSIONS RESULTING FROM THE EXAMINATION AND TESTING THE RESEARCH QUESTIONS AND HYPOTHESES

First research question: How much the managers possess the managerial skills of a manager?

Results showed that technical skills had the highest score (mean 3.61, SD 0.32) and conceptual skills had the lowest score (mean 3.37, SD 0.364). So, considering the obtained results, it can be concluded that most of the staff believe that managers have high technical skills and low conceptual skills. This result is consistent with the research of Mojtaba Hosseini Nia (2007). From the results of his study, he concluded that technical skills are the most important factor in an organization's effectiveness. This factor is evaluated in his study at a high level.

The above results are consistence with the study of Johnson (2000), in his opinion the important factor in the effectiveness of manager on managing the staff is his area of expertise and skill (technical skills).

Research Question 2: How much the managers possess managerial competence?

The results showed that the subjects of the study in case of their competence, academic competence had the mean of 3.48 and SD of 0.375 which was the lowest mean. And the moral competence with the mean of 3.55 and SD of 0.351 had the highest mean. So, as the data presented in the aforesaid table shows, it can be concluded that in the opinion of the employees, the managers of the Ilam Telecom Company have high moral competence and low academic competence. The categorization of the competencies from the highest to the lowest competence is as follows.

Moral competence > Managerial Competence > Conceptual Competence > Academic competence

To conclusion of the current study is in line with the research of Ahmadi (2003), in his study there was a significant relationship between gender and the level of human skills, conceptual and technical.

The results of the foresaid question was consistent with Rabbani's research results (1380), he concluded that there is a significant relationship between gender and the three competencies.

The third hypothesis of the research: is there a significant relation between the academic field of the study of the staff and their assessment of the managers' managerial skills?
The findings showed that a significant level of 0.05 and the confidence level of 95%:

- There is no significant relationship between the academic field of study and the conceptual and human skill.
- There is a significant relation between the academic field of study and technical skills.

The result of this study is the same as the research results of Mir sadooghi (2006). In his study he found that the assessment of the human skill of the managers by the staff is not different based on whether they have a management degree or not. And statistically there is no significant relation between the two groups. However, the mean of the human skills of the managers with a degree other than management was higher than the managers who had a degree in management.

The fourth hypothesis of the study: is there significant relation between the academic level of employees and their assessment of the three managerial skills?

The results showed that in a significant level of 0.05 and the confidence level of 95% there is a significant difference between the technical and conceptual skills, but this is not true for human skills.

The above hypothesis regarding the human skill was in line with the research of Sheikhzadeh(2008), he believed that there is no significant relation between the level of education and the level of possessing the human skills. The results of the hypothesis regarding the human skills contrasts the results of Pezeshk (1999), but regarding the two other skills the results is the same.

REFERENCES

[1] Ahmadi , A. (2003). The rate of secondary school principals in the city Kamyaranplainis skills (technical, human and conceptual) and their relationship with the effectiveness of teachers' views of the school year 2002-2003 , MS Thesis martyr Beheshti University in Educational Administration

[2] AlagheBand, A.(2005).Public Management(Second Edition),Tehran, publishmental.

[3] Drucker, P. (2003). No son empleados, son personas (Thay are not employees, thy are people), Trend management, 4 (4): 16-22.

[4] Hammer, J & champ, R .(2005). Skill of effective administration,Harvard business review,vol

[5] Hosseini -Nia , M. (2007). The survey of executives modernization, development and equipping of schools from Triple Management Skills (technical, human, conceptual) looking at the impact on employees and their effectiveness , Tehran , master's thesis martyr Beheshti University in Educational Administration

[6] Johnson,S. (2000). The concept of Competency: An Operational Defination.Educational Technology,18 (2).

[7] Khaki, G.(2011).Approachto theresearchdissertation.Tehran, publishingreflection

[8] Katzenbach, G & smith, E. (2003). Principles of modern managerial skills.IIIinons peacock publisher Inc.

[9] Mo'ayyeri,M.T.(1998).Education issues, Eighth Edition,Tehran,Amir KabirPublications7-

[10] Myrsdvqhy , S.H, (2006). Study of human skills in Qazvin school administrators and their relationship with burnout, Journal of Knowledge Management, No. 3 , Summer 84.pp.132-149

[11] Pezeshk , A.S, (1999). City high school teachers opinions about the characteristics of cognitive, emotional , and functional requirements for effective managers in high school , " Master's Thesis in Educational Administration University

[12] Shenhar & Reiner.(2001). Leadership Skill for manager, New york: By McGraw-Hill

[13] Viitala, R. (2005). Perceived development needs of managers compared to an integrated management competency model, Journal of Workplace Learning, 17 (7): 36-41.

[14] Sheikhzadeh , M. (2008). Comprehensive look at the managers' performance evaluation model , Tehran Electronic Journals research group wide strategy , first issue , Fall 86

M-HEALTH AN EMERGING TREND AN EMPIRICAL STUDY

Muhammed Fuzail Zubair, Hajrah Jahan, Dr. Sophia Rahaman and Dr. Roma Raina

School of Engineering and IT, Manipal University, Dubai, UAE

fuzailmuhammed@gmail.com
hajrah.jahan@yahoo.com
sophia@manipaldubai.com
roma.raina@manipaldubai.com

ABSTRACT

The advent and advancement in technology specific to medical field has seen a migration of its work across the globe, adapting higher and newer levels of m-health. Technology has been successful in transforming the way traditional monitoring and alert system work to a modern approach wherein minimizing the need for physical monitoring. Today, the field of healthcare use varied monitoring systems to monitor the health of patients using ubiquitous and non-ubiquitous devices. These are sensor based devices that can read vital signs of patients and send the data to the required personnel's using mobile networks. This paper understands and analyses how the monitoring and alert system works specific to m-health. m-health including wearable and non-wearable devices read various vital signs and have the ability to monitor health real-time and transfer the information collected using mobile network. m-health has become an useful tool for elderly in this fast paced world where almost all the family members are working or studying to keep track and maintain optimal health status. m-health alert system involves the patient, the caretaker and medical service provider wherein the patient wears the device and vital signs recorded are transferred the medical service provider who then analyses the data collected and required changes in the medication are implemented. This paper proposes a medical alert system that enlightens the capabilities of m-health making health monitoring easy and reliable. It contains a three-level severity check and raises an alarm to the caretaker, the physician or the ambulatory service provider.

KEYWORDS

e-health; m-health; Wearable Wireless Sensor based Area Network; Data Mining.

1. INTRODUCTION

The advent of mobile technology and its services are quickly emerging as the new frontier in transforming the public and private sectors of the society, trying to make it even more accessible and customer-centric by extending the benefits of remote delivery [1]. Their main objective is to provide a 24x7 service across the globe via mobile networks. Particularly in the field of medicine

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 167–174, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60215

the approach has elevated healthcare practice to a new level that imbibes computing power backed by the electronic communication technology [2].This approach spans across various services like, Electronic health records, Computerized physician order entry, e-prescribing, Clinical Decision Support, Telemedicine, Consumer health informatics, Health knowledge management, Medical research using Grids [3] & Healthcare Information Systems.

e-health is categorized as ubiquitous where the monitored health resides on a particular location and is communicated to the interested personal via direct connection to the server. For example, World Health Organization's consumer health information service. On the other hand is non-ubiquitous where the communication uses wireless technology in communicating the information to the user, e.g. CH Telemetry System by ApexPro. Latest advancements are related to another term called m-Health which is monitoring and communicating medical conditions using mobile technology [4,5]. The fast pace adaption to m-health is due to the major shift in health care needs towards more scalable and affordable solutions. This is restructuring the healthcare systems towards proactive management of wellness rather than illness, which is focused at prevention and early detection of disease [6].

2. MOTIVATION AND OBJECTIVE

The classification of m-health has routed to a path of non-wearable and wearable devices. The non-wearable technology in m-health constitutes of applications to sense the heartbeat, blood pressure, oxygen saturation, UV exposure, pedometer etc. and is widely used by the leading mobile manufacturers like Samsung & Apple. These technologies require a head on to start their process and provide an output. On the contrary, wearable technologies focus is to provide a real time health monitoring service to maintain optimal health status. When integrated into a telemedical system they raise a medical alert in life threatening scenarios. In addition, patients also benefit in terms of continuous long term monitoring as a part of a diagnostic procedure to achieve optimal maintenance of a chronic condition, or supervision during recovery from an acute event or surgical procedure.

Wearable health monitoring systems [7] integrated into a telemedicine system is a novel information technology approach that will be able to support early detection of abnormal conditions and prevention of its serious consequences [8]. The wearable devices shelter numerous sensors required for monitoring health conditions like, pulse, blood pressure, temperature and others [9]. These devices worn or carried by the patient, monitor the condition on a regular basis and provide real time information to the receiving end of the communicating device, particularly mobile phones in this context, via applications, IM, links etc. Keeping in view that health monitoring has the potential to improve the quality of health services & ensuring that those who need urgent care get it sooner, this research aims to study and understand the varied dimensions of such systems [10,11,12].

3. METHODOLOGY

Recent technology advances in integration and miniaturization of physical sensors, embedded microcontrollers and radio interfaces on a single chip; wireless networking; and micro-fabrication have enabled a new generation of wireless sensor networks suitable for such applications by the use of Wearable Wireless Sensor based Area Network (WWSAN) [13,14]. The WWSAN consisting of inexpensive, lightweight, and miniature sensors carry long-term, unobtrusive,

ambulatory health monitoring with instantaneous feedback to the user about the current health status and real-time or near real-time update of the user's medical records. Intelligent heart monitors warning users about impeding medical conditions[15],Accelerometer-based system monitoring physical activity [16] are a few to name.

The wearable device can include a number of physiological sensors and photoplethysmographic biosensors [17] depending on the end-user application. An extensive set of physiological sensors may include an (electrocardiogram) sensor for monitoring heart activity, an EMG (electromyography) sensor for monitoring muscle activity, a blood pressure sensor, a temperature sensor, an EEG (electroencephalography) sensor for monitoring electrical activity of the brain etc.

The WWSAN when integrated into a broader telemedical system with patients' medical records promises a revolution in medical research with the use of data mining on the gathered information. This enables the researchers to explore the synergy between varied parameters of the collated data giving scope to perform quantitative analysis of various conditions and patterns that will built be with time. The Sensors on the wearable device can be integrated into various objects such as garments, wrist bands, socks, shoes etc. Some of the sensors that are invasive in nature are implanted on to the body. The sensor plays a significant role by collecting various data easily through signals on a chip that is ingested or worn by the patient body and helps in storing the various changes that occur in an organized form on a device the sensor is connected to and provide an alert to immediate caretakers and other emergency contacts based on the severity of variation as depicted in figure.

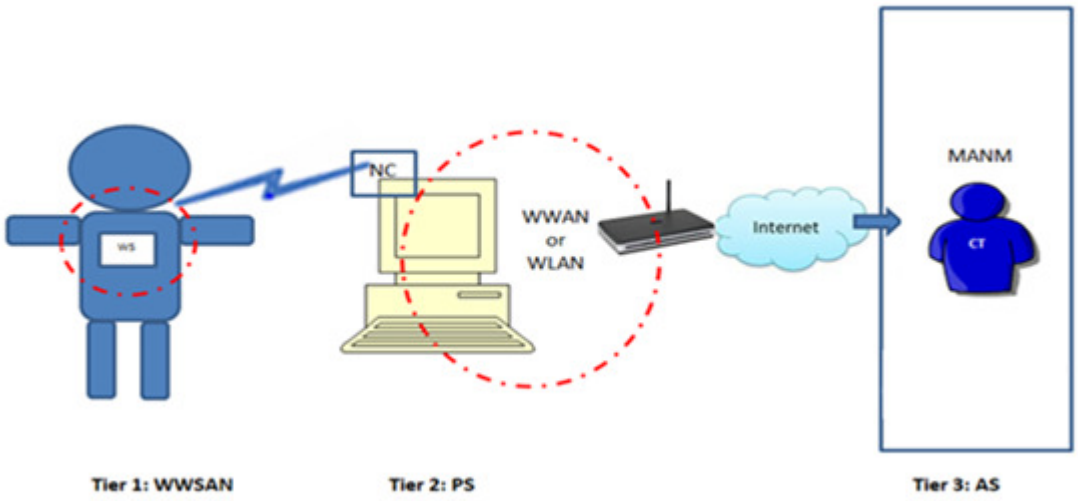


Figure 1: Three Tier Medical

As depicted in Figure 1, the first tier is patient wearing the sensor (WS), via the sensor nodes, it senses and processes the varied physiological signals. The WWSAN capture these signals and via Network Configuration (NC) transmit them to tier 2 that hosts the personal server (PS). The PS provides transparent interfaces to the WS; the device then uses the WWAN or WLAN and via the internet raises a medical alert notification message (MANM). Finally PS communicates the medical condition to the 3rd tier, Alert System(AS) which raises a MANM based on severity of the patient's condition either to the care taker (CT) or an emergency contact apart from updating

the patient medical record. This technology can improve the quality of life of elderly and dependent people who require medical attention frequently and immediately [18].

4. DISCUSSION

The proposed methodology in the previous section is further explained with the use of case studies. Two case studies are chosen to illustrate working of WS in a health monitoring system provided to the target group over the pilot period of the experiment.

First case study is of CHRISTUS Health System [19] by St. Michael Hospital, Texarkana. They used CHRISTUS Health System to monitor the health of high-risk patients diagnosed with chronic illness such as Congestive Heart Failure, Hypertension, and Diabetes. This health system consists of Remote Patient Management System, Remote Patient Monitoring System that includes Android tablet and Bluetooth paired personal health devices: weight scale, blood pressure monitor, pulse oximeter. In this health system, patient protocols and care plans are easily customized for each patient and an user interface is provided for the patients to use. At the end of the program, results showed a 90% reduction in overall cost of care, 65% reduction in hospital readmissions, 95% patient adoption and patient satisfaction.

Second case study is of Ambio Health Remote Patient Monitoring conducted by Enjoy Life! [20] Health Consulting. Diabetic patients were invited to participate in a pilot program in which they were given blood glucose meter, blood glucose strips and supporting equipment to send their blood glucose reading wirelessly. Blood pressure monitor was also provided based on the requirement. A patient who volunteered for this program had a history type 2 diabetes and high blood pressure. The patient never used to check blood pressure and would forget the blood sugar levels or doctor found the blood sugar levels provided by him were hundred percent correct. After using this system, his health care provider found out that all the numbers were high based on the data collected and prompt action was taken with change in medication in consultation with the patient's physician.

The underlying application of data mining enables the learners to digest large amounts of data by leveraging sophisticated techniques in data analysis, restructuring and organization [21]. In a datamining system the data is preprocessed and suitable mining technique is applied on the relevant data that are generally descriptive and predictive in nature and generate patterns as knowledge to the user. Some of the commonly used techniques for predictive analysis specific to healthcare are as discussed in brief :

Clustering - This approach is the stratification of objects into different groups, precisely partitioning the data set into clusters or subsets. The subset ideally shares certain common straits. This technique is suitable for statistical data analysis used in fields like machine learning, pattern recognition, image analysis and bioinformatics.[22]

Classification – This approach is known for its categorization of data for efficient and effective use. Proper classification of patient health records aid in verification, diagnosis and in-depth data processing. Classification tree and rules are commonly used to obtain simple models that are employed in clinics. Recent examples of such models include classification of cancer using gene expression data, classification of tumors of the tongue to name a few. [23,24]

Association rule mining – This approach enables the users in finding interesting patterns and trends in data. Association rules identify collections of data attributes that are statistically related in the underlying data. This approach brings out the precautionary measures that can be adapted for better health management. [25, 26, 27]

Sequential mining – This technique in specific is focused in finding inter-transaction patterns, to detect the presence of a set of events in a time ordered sequence of transactions. Here an order exists between the occurring events with a possibility of an event re-occurring in the same order. This functionality makes it aptly suitable to the healthcare sector in predicting future health risks and educating in better health management.[28, 29].

5. PROPOSED MODEL

An analysis of the cases discussed in the above section brings to light the importance of remote health monitoring system. It emphasizes on the availability, ease of use, readiness and significance of such systems. The proposed model depicted in figure 2, illustrates the process flow for these systems.

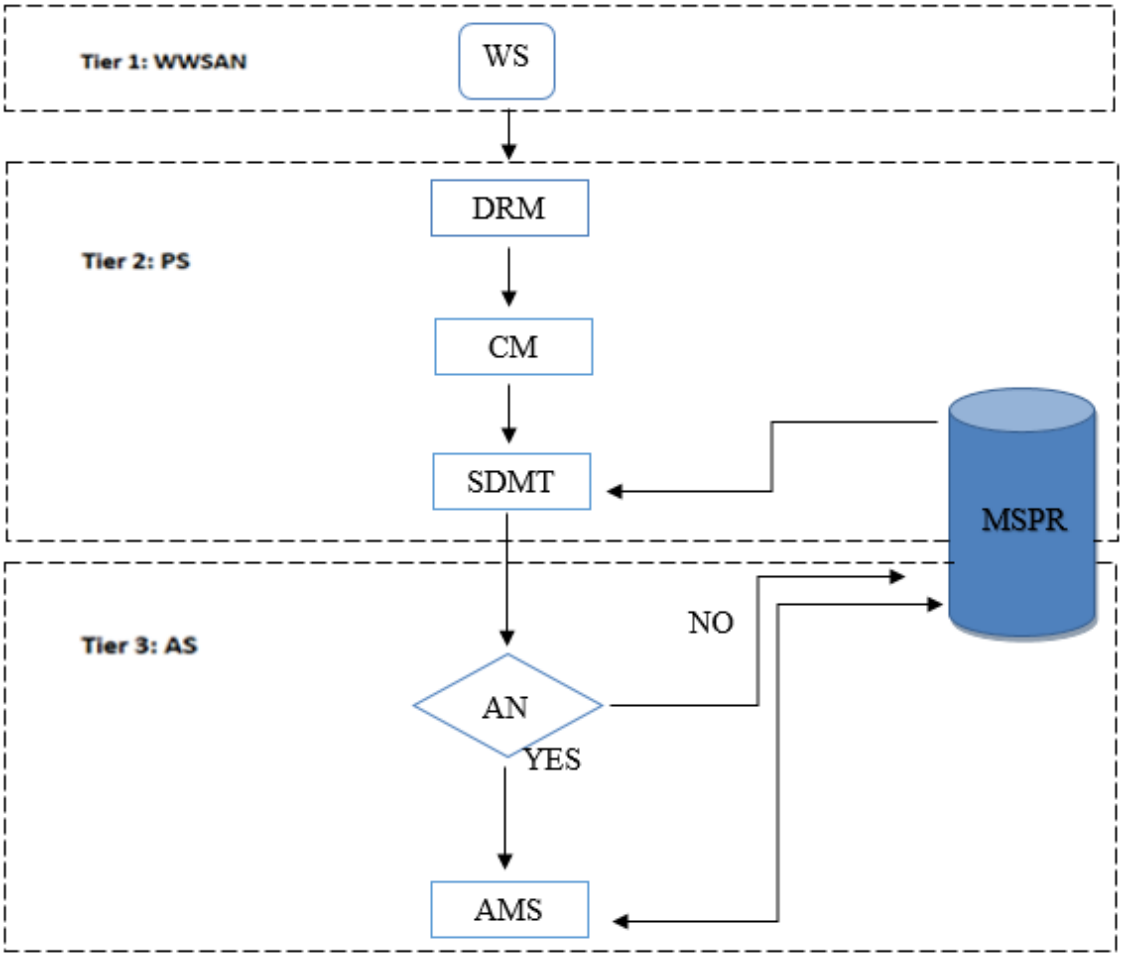


Figure 2: Process flow of the proposed model

The proposed model states the wearable sensor (WS) senses the required vital readings and transmits them to the Data Receiver Module (DRM) in the second tier: PS. The DRM processes the data as per the measurement of the system and passes it to the Compilation Module (CM). The compiled data in now further analysed by the application of Suitable Data Mining Techniques (SDMT). The analysis results are passed to the Alert Notification Module (ANM) in tier 3: AS. The AVM raises an alert based on the severity of the patient's condition to the notification receiver (emergency contact). The Ambulatory Service (AMS) and the Medical Service Provider (MSP) get the access to the PRMH along with the Condition Prevalent Data (CPD).

6. CONCLUSION AND FURTHER RESEARCH

This paper demonstrates the use of WWSAN as a key infrastructure enabling unobtrusive, continual, ambulatory health monitoring. This emerging technology offers wide range of benefits to patients, medical personnel, and the society through continuous monitoring in the ambulatory setting, early detection of abnormal conditions, supervised rehabilitation, and potential knowledge discovery through data mining from the gathered information.

The increasing development of mobile health care system yields the largest growth among mobile users. Mobile healthcare alert system that delivers the proper timing and emergency case alerts is considered advantageous related to power consuming, portability and flexibility as the mobility devices enhance the computation based on the ubiquitous nature. In addition, further studies of varied medical conditions in clinical and ambulatory settings are essential to determine specific limitations and possible new and applications of this emerging technology.

REFERENCES

[1] Park. S, (2003) "Enhancing The Quality Of Life Through Wearable Technology", Engineering In Medicine And Biolog Magazine, IEEE, Volume:22 Issue: 3.

[2] Della Mea & Vincenzo (2001) "What Is E-Health (2): The Death Of Telemedicine?", Journal Of Medical Internet Research 3.

[3] Jochen Fingberg & Marit Hansen Et Al., (2006)"Integrating Data Custodians In Ehealth Grids – Security And Privacy Aspects", NEC Lab Report.

[4] Adibi & Sasan, (2015) "Mobile Health: A Technology Road Map. Springer", Ed. February 19, Isbn 978-3-319-12817-7.

[5] K. Kiran Reddy, P.Lalith Samanth Reddy & Dr.P.Bhaskara Reddy, (2014) "Study On Mobile Healthcare System", IJARCSSE, Volume 4, Issue 3, March. Available At: Www.Ijarcsse.Com

[6] Aleksandar Milenković, Chris Otto & Emil Jovanov, (2006) "Wireless Sensor Networks For Personal Health Monitoring: Issues And An Implementation".

[7] Rutherford & J.J, (2010) "Wearable Technology", Engineering In Medicine And Biolog Magazine, IEEE, Volume:29 Issue: 3.

[8] Istepanian RSH, Jovanov E & Zhang YT, (2004) Guest Editorial Introduction To The Special Section On M-Health: "Beyond Seamless Mobility And Global Wireless Health-Care Connectivity" IEEE Transactions On Information Technology In Biomedicine, 8(4):405-414,Pubmed Abstract.

[9] Dr. Rajender Thusu, (2011) "Medical Sensors Facilitate Health Monitoring",Frost & Sullivansensors, 1st April. Available At: Http://Www.Sensorsmag.Com/Specialty-Markets/Medical/Sensors-Facilitate-Health-Monitoring-8365

[10] Bonato. P , (2010) "Wearable Sensors And Systems", Engineering In Medicine And Biolog Magazine, IEEE, Volume:29 Issue: 3.

[11] Ching Hsu, Member & IACSIT, (2013) "Wireless Context-Aware Healthcare System Based On Sensor", Web 2.0 I-, IJI, Management And Technology, Vol. 4, No. 4, August.

[12] Jessica Gomez, (2014) "25 Most Interesting Medical MEMS And Sensor Projects", Rogue Valley Microdevices. Available At: Http://Www.Slideshare.Net/Mikepinelisphd/140804-25-Most-Interesting-Medical-Mems-Sensors

[13] E. Jovanov, A. Milenkovic, C. Otto & P. C. De Groen, (2005) "A Wireless Body Area Network Of Intelligent Motion Sensors For Computer Assisted Physical Rehabilitation," Journal Of Milenkovic Et. Al. Neuroengineering And Rehabilitation, 2:6, 1st March. Available At: [Http://Www.Jneuroengrehab.Com/Content/2/1/6]

[14] Jeonggil Ko, Chenyang Lu, Mani B. Srivastava, John A. Stankovic, Andreas Terzis & Matt Welsh, (2010) "Wireless Sensor Networks For Healthcare", Proceedings Of The IEEE, Vol. 98, No. 11, November.

[15] J. Welch, F. Guilak & S.D. Baker, (2004) "A Wireless ECG Smart Sensor For Broad Application In Life Threatening Event Detection", In Proceedings Of The 26th Annual International Conference Of The IEEE Engineering In Medicine And Biology Society, (San Francisco, CA, September), Pp. 3447-3449.

[16] M.J. Mathie & B.G. Celler, (2001) "A System For Monitoring Posture And Physical Activity Using Accelerometers", In Proceedings Of The 23rd Annual International Conference Of The IEEE Engineering In Medicine And Biology Society, 2001, Pp. 3654- 3657.

[17] Asada, H.H, (2003) "Mobile Monitoring With Wearable Photoplethysmographic Biosensors", Engineering In Medicine And Biology Magazine, IEEE, Volume:22 Issue: 3.

[18] V.Barath & P.Anithapriya, (2013) "A Survey On Healthcare For Aged People Based On Sensor Network", IJIRCCE, Volume:1 Issue:8, October.

[19] Christus Health System, St. Michael Hospital, (2014) "Remote Patient Monitoring for Care Transition Intervention Program, utilizing Remote Patient Monitoring System (RPMS) from Vivify Health".

[20] Leading Age Cast – Centre for aging services technologies, (2013) "Telehealth And Remote Patient Monitoring (RPM)".

[21] Guangming Li, (2013) "Research On The Medical Sub- Health Status Prediction And Future Selection", IJACT Vol 5, No. 2, Pp 226~232.

[22] Khaled Hammoudal & Mohammed Kamel, (2005) "Data Mining In E- Learning", Pattern Analysis And Machine Intelligence (PAMI) Research Group, University Of Waterloo, Canada, 19 July.

[23] Chan Sheung Wai & Clement Leung, (2013) " Clinical Decision Suport Systems For Acute Leukemia Classification In Co-Developments", JDCTA, Vol 7, No. 1, Pp232 ~ 239.

[24] Fabricio Voznika & Leonardo Viana, (2007) "Datamining Classification".

[25] C.S.Kanimozhi Selvi & A.Tamilarasi, (2009) "An Automated Association Rule Mining Technique With Cumulative Support Thresholds" , Int. J. Open Problems In Compt. Math, Vol. 2, No. 3, September, ISSN 1998-6262; Copyright © ICSRS Publication. Available At: Www.I-Csrs.Org

[26] Gary S. Firestein, Ralph C. Budd, Sherine E. Gabriel, Iain Mclnnes & James R O'Dell, (2000) "Kelly's Textbook Of Rheumatology", WB Saunders Co., Edited By Shuan Ruddy.

[27] Lippincott Williams & Wilkens, (2003) "Clinical Primer Rheumatology", Edited By William Koopman Et, Al, Association.

[28] Ji˘r´I Klema, Lenka Nov ´ Akov ´ A, Filip Karel, Olga ´ St˘ Ep˘ Ankov ´ A & Filip ´ Zelezn ˘ Y, (2008) "Sequential Data Mining: A Comparative Case Study In Development Of Atherosclerosis Risk Factors", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 38, No. 1, January.

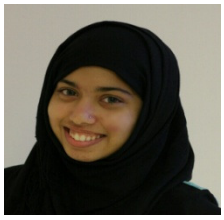
[29] Sophia Banu Rahaman & Prof. M. Shashi, (2011) "Sequential Mining Equips E-Health With Knowledge For Managing Diabetes", IJIPM, Vol. 2, No. 3, Pp. 74 ~ 85.

AUTHORS

Muhammed Fuzail Zubair

Author is currently pursuing undergraduate degree in B.Sc Information System and Management with Networks as his specialization. His interest lies in Internet of Things (IOT), Cloud Technologies and Network Technologies. Excellent programmer with knowledge in networks and IT hardware, keen in taking up interesting and challenging projects.

Hajrah Jahan



Author is currently pursuing undergraduate degree in B.Sc Information System and Management with Software as her specialization. Her interest lies in developing applications using various programming languages. An ambitious programmer, keen in taking up interesting projects and an aspiring writer.

Dr. Sophia Rahaman

Author currently works as Professor in Manipal University,Dubai, imparts knowledge to students of Information Technology and Computer Science Engineering. Her area of interest lies in Databases, Data Mining and Digital World. Her area of expertise lies in Approaches and Technologies of Sequential Mining, E Health and E Learning. Author has published several research papers in her area of expertise

Dr. Roma Raina

Author currently works as Professor in Manipal University, Dubai, imparts knowledge to students of Engineering. Her area of interest lies in Controls & Instrumentation. Her area of expertise lies in Radical Power Distribution, Controls & Intstrumentation, Digital Electronics and Electrical Machines. Author has published several research papers in her area of expertise.

A COMPREHENSIVE SURVEY OF LINK MINING AND ANOMALIES DETECTION

Dr. Zakea Idris Ali

Staffordshire University- UK
zakiazi@hotmail.com

This survey introduces the emergence of link mining and its relevant application to detect anomalies which can include events that are unusual, out of the ordinary or rare, unexpected behaviour, or outliers.

1. EMERGENCE OF LINK MINING

Link mining is a newly developed research area, bringing together research insights from the fields of web mining, graph theory and machine learning. Link mining applications have been shown to be highly effective in addressing many important business issues such as telephone fraud detection (Fawcett & Provost, 1999), crime detection (Sparrow, 1991), money laundering (Kirkland *et al*., 1999), terrorism (Badia & Kantardzic, 2005; Skillicorn, 2004), financial applications (Creamer & Stolfo, 2009), social networks and health care problems (Provana *et al.,* 2010; Wadhah *et al*., 2012). The trend in the building and use of link mining models for critical business, law enforcement and scientific decision support applications are expected to grow. An important issue will be building models and techniques that are scalable and reliable.

Link mining attempts to build predictive or descriptive models of the linked data (Getoor & Diehl, 2005). The term 'link' in the database community differs from that in the AI community. In this research a link refers to some real-world connection between two entities (Senator, 2005). Link mining focuses on techniques that explicitly consider these links when building predictive or descriptive models of the data sets (Getoor, 2005). In data mining, the main challenge is to tackle the problem of mining richly structured heterogeneous data sets. The data domains often consist of a variety of object types; these objects can be linked in a variety of ways. Traditional statistical inference procedures assume that instances are independent and this can lead to unsuitable conclusions about the data. However, in link mining, object linkage is a knowledge that should be exploited. In many applications, the facts to be analysed are dynamic, so it is important to develop incremental link mining algorithms, besides mining knowledge from link objects and networks (Getoor & Diehl, 2005).

2. LINK MINING TASKS

In their paper, Getoor and Diehl (2005) identify a set of link mining tasks (see Figure 1), which are:

- Object-related tasks.
- Graph-related tasks.
- Link-related tasks.

2.1 Object-related tasks

These tasks include link-based object clustering, link-based object classification, object identification and object ranking. In a bibliographic domain, the objects include papers, authors,

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 175–189, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60216

institutions, journals and conferences. Links include the paper citations, authorship and co-authorship, affiliations, and the relation between a paper and a journal or conference.

2.2 Graph-related tasks

These tasks consist of sub-graph discovery, graph classification, and generative models for graphs. The aim is to cluster the nodes in the graph into groups sharing common characteristics. In the bibliographic domain, an example of graph classification is predicting the category of a paper, from its citations, the papers that cite it, and co-citations (papers that are cited with this paper).

2.3 Link-related tasks

These tasks aim at predicting the existence of a link between two entities based on the attributes of the objects and other observed links. In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.

Link prediction is defined as inferring the existence of a link (relationship) in the graph that is not previously known. Examples include predicting links among actors in social networks, such as predicting friendships or predicting the participation of actors in events (O'Madadhain et al., 2005) such as email, telephone calls and co-authorship. Some links can be observed, but one is attempting to predict unobserved links, or monitor the temporal aspect; for example, if a snapshot of the set of links at time t is observed then the goal is to predict the links at time t + 1.

This problem is normally expressed in terms of a simple binary classification problem. Given two potentially linked objects Oi and Oj, the task is to predict whether Lij is 1 or 0. One approach bases the prediction on the structural properties of the network, for example using predictors based on different graph proximity measures Liben-Nowell and Kleinberg (2003). The second approach is to use attribute information to predict a link. Popescul et al. (2003) applied a structured logistic regression model on relational features to predict the existence of links. A conditional probability model is proposed which is based on attribute and structural features by O'Madadhain et al (2005); (Getoor, 2003; O'Madadhain, 2005; Rattigan & Jensen, 2005). They explain that building statistical models for edge prediction is a challenging problem because the prior probability of a link can be quite small, this makes it difficult to evaluate the model and, more importantly, measure the level of confidence in the predictions. Rattigan and Jensen (2005) propose improving the quality of the predictions by making the predictions collectively. Hence, a number of probabilistic approaches have been developed, some network structure models are based on the Markov Random Field (MRF) model (Chellappa & Jain, 1993) others on Relational Markov Network (Taskar et al., 2003) and, more recently, the Markov Logic Network (Domingos & Richardson, 2004). If case, O represents a set of objects, with X attributes, and E edges among the objects, then MRF uses a joint distribution over the set of edges E, P(E), or a distribution conditioned on the attributes of the nodes, P(E/X). Getoor et al (2003) described several approaches for handling link uncertainty in probabilistic relational models. The key feature of these approaches is their ability to perform probabilistic inferences about the links, which allows the capture of the correlations among the links. This approach is also used for other tasks, such as link-based classification, which allow for more accurate predictions. Hence, approximate inference techniques are necessary to join the model-based probabilistic approaches based on their computational cost to exact inference as general intractable goals.

Desjardins and Gaston (2006) discuss the relationship between the fields of statistical relational learning (SRL) and multi-agent systems (MAS) using link prediction methods to recognise collusion among agents, and applying graph classification to discover efficient networks for MAS

problems. Mustafa et al. (2007) show a general approach for combining object classification and link prediction using Iterative Collective Classification and Link Prediction (ICCLP) in graphs.

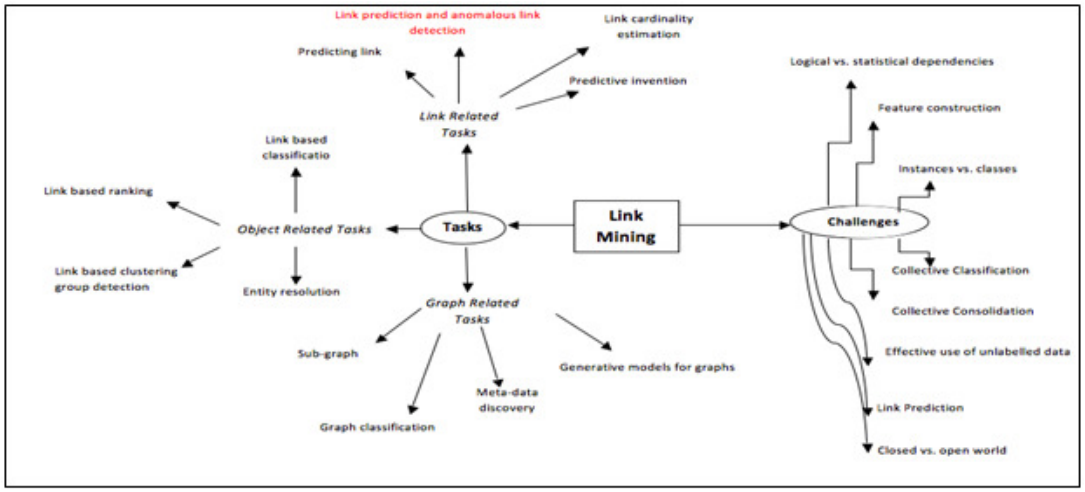


Figure 1. Link mining tasks and challenges

3. LINK MINING CHALLENGES

Research into link mining involves a set of challenges associated with these tasks, as Senator (2005), Getoor (2005) and Pedreschi (2008) explain (see Figure 1). These are:

- logical vs statistical dependencies that relate to the identification of logical relationships between objects and statistical relationships between the attributes of objects;
- feature construction, which refers to the potential use of the attributes of linked objects;
- collective classification using a learned link-based model that specifies a distribution over link and content attributes, which may be correlated through these links;
- effective use of unlabelled data using semi-supervised learning, co-training and transductive inference to improve classification performance;
- link prediction, which predicts the existence of links between objects;
- object identity, that is, determining whether two objects refer to the same entity; and closed world vs open world assumptions of whether we know all the potential entities in the domain.
- the challenge of this study is to identify and interpret anomalies among the observed links.

4. APPLICATIONS OF LINK MINING

An application for each of the three tasks is listed below.

- Social bookmarking is an application of a link-related task. Tools enable users to save URLs for upcoming reference, to create labels for annotating web pages, and to share web pages they found interesting with others. The application of link mining to social web bookmarking

investigates user bookmarking and tagging behaviours, and describes several approaches to finding patterns in the data (Chen & Pang-Ning, 2008).

- Epidemiological studies are an application associated with object-related task. In an epidemiology domain, the objects include patients, people with whom they have come into contact and disease strains. Links represent contacts between people and a disease strain with which a person is infected (Getoor, 2003).
- Friendship in a social network is an application of graph-related task. This is annotated by the inclusion of the friend's name on a user's homepage. Pair-dependent descriptions, such as the size of the intersection of interests, offer supplementary evidence for the existence of a friendship. These pair-dependent features are used to determine the probability for link existence where it is not annotated. Finding the non-obvious pair-dependent features can be quite difficult as it, requires the use of recent developments in association rule mining and frequent pattern mining to find correlations between data points that best suggest link existence (Han *et al*., 2001).
- Bibliographic area is an application of a graph-related task. Information networks are mainly new. Link information in a bibliographic database provides in-depth information about research, such as the clustering of conferences shared by many common authors, the reputation of a conference for its productive authors, research evolving with time, and the profile of a conference, an author, or a research area. This motivates the study of information network in link mining on bibliographic databases (Getoor, 2003).
- Discovery of a fundamental organisation is an application of graph-related task. Structure from crime data leads the investigation to terrorist cells or organised crime groups, detecting covert networks that are important to crime investigation. (Marcus *et al*., 2007).

5. ANOMALIES DETECTION

Link prediction is a complex and challenging task as many applications contain data which are extremely noisy and often the characteristics to be employed for prediction are either not readily available or involve complex relationships among objects. The focus of this thesis is to investigate the links between objects and understand the context of their anomalies. Anomaly detection is different from noisy data, which is not of interest to the analyst, and must be removed before any data analysis can be performed. In our research anomalous objects or links can convey useful information and should be investigated.

Song *et al*. (2007) and Chandola *et al.* (2009) describe five types of anomalies, these are:

- Contextual anomalies (also known as conditional anomalies) refer to data instances anomalous in a specific context. A temperature of $5^{o}C$ might be normal during the winter period in the UK, but would be an anomaly in the summer time.

- Point anomalies refer to a data instance anomalous with respect to the rest of the data set. In credit card fraud application, a transaction is considered a point anomaly if it contains a very high amount spent compared to the normal range of expenditure for that individual.

- Collective anomalies refer to a set of data instances anomalous with respect to the entire data set. For example an electrocardiogram output may show a region of low values for an abnormally long time due to some premature contractions (Goldberger *et al.,* 2000). These low values may not be anomalies by themselves, but their existence together as a collection is anomalous.

- On-line anomalies refer to data present often in a streaming mode where the normal behaviour is changing dynamically.

- Distributed anomalies refer to detecting anomalies in complex systems.

The definition of anomaly is dependent on the type of application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) could be an anomaly, however similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another has to take into consideration the context of that domain.

Anomalies detection is alike to link prediction in the sense that they both use similar metrics to evaluate which links are anomalous and which ones are expected. Thus research on improving either problem should benefit the other. Rattigan and Jensen explain that one of the important challenges in link prediction is to address the problem of a highly skewed class distribution caused by the fact that " ... as networks grow and evolve, the number of negative examples (disconnected pairs of objects) increases quadratically while the number of positive examples often grows only linearly" (Rattigan and Jenssen 2005: 41). As a result, evaluating a link prediction model becomes a complex task and computationally costly because of the need to evaluate all potential links between all pairs of objects. They have proposed the alternative task of anomalous link discovery (ALD) focusing on those links that are anomalous, statistically unlikely, and most "interesting" links in the data. Typical applications of anomaly detection algorithms are employed in domains that deal with security and privacy issues, terrorism activities, picking intrusion detection and illegitimate financial transactions (See Figure 1).

6. ANOMALIES DETECTION APPROACHES AND METHODS

A survey of the literature reveals three main approaches used to detect anomalies. These are described below:

- *Supervised* anomalies detection operates in supervised mode and assumes the availability of a training data set, which has labels available for both normal and anomalous data. Typical approach in such cases is to build a predictive model for normal vs. anomalous classes; their disadvantage is that they require labels for both normal and anomalous behaviour. Certain techniques insert artificial anomalies in a normal data set to obtain a fully labelled training data set and then apply supervised anomalies detection techniques to detect anomalies in test data (Abe *et al*., 2006).

- *Semi-supervised* anomalies detection, which models only normality and are more applicable than the previous approach since only labels for normal data is required. Such techniques are not used commonly, as it is difficult to obtain a training data set, which covers possible outlying behaviour that can occur in the data (Chandola *et al.,* 2009).

- *Unsupervised* anomalies detection, which makes the implicit assumption that normal instances are more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from a high false alarm rate (Chandola *et al.,* 2009).

Unsupervised method is very useful for two reasons. First, they do not rely on the availability of expensive and difficult to obtain data labels; second, they do not assume any specific characteristics of the anomalies. In many cases, it is important to detect unexpected or unexplained behaviour that cannot be pre-specified. Since the unsupervised approach relies on

detecting any observation that deviates from the normal data cases, it is not restricted to any particular type of anomaly.

In their paper, Chandola *et al. (*2009) identify five different methods employed in anomalies detection: nearest neighbour, clustering, statistical, classification, and information/ context based approaches (see Figure 2).

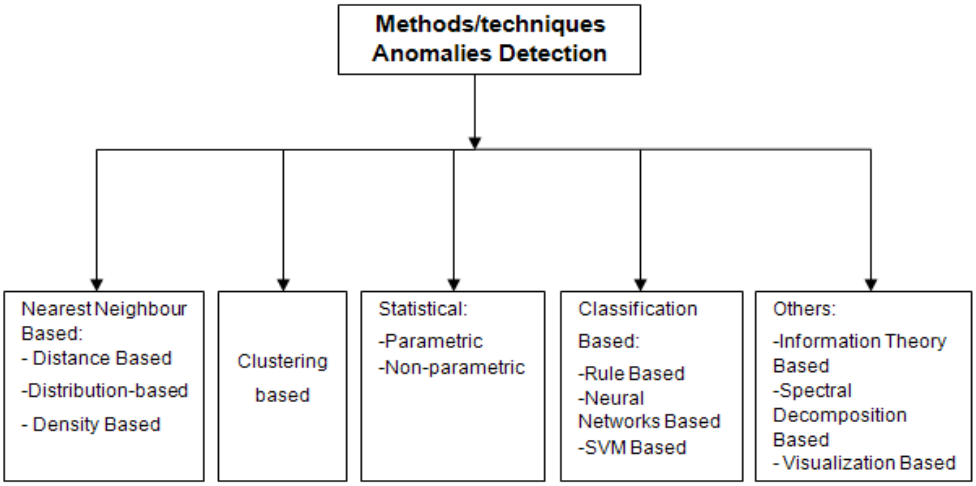


Figure 2. Methods of anomalies detection

6.1 Nearest neighbour based detection techniques

The concept of nearest neighbour has been used in several anomaly detection techniques. Such techniques are based on the following key assumption:

Assumption*:* Normal data instances happen in dense neighbourhoods, while anomalies occur far from their closest neighbours.

The nearest neighbour based method can be divided into three main categories. The first distance-based methods, distinguish potential anomalies from others based on the number of objects in the neighbourhood (Hu and Sung, 2003). The distribution-based approach deals with statistical methods that are based on the probabilistic data model, which can be either a automatically or priori, created using given data. If the object does not suit the probabilistic model, it is considered to be an outlier (Petrovskiy, 2003). The density-based approach detects local anomalies based on the local density of an object's neighbourhood (Jin *et al*., 2001). A typical application area is fraud detection (Ertoz *et al.*, 2004; Chandola et al. 2006), Eskin *et al* (2002).

Nearest neighbour based techniques have many advantages. Key advantage is that they are unsupervised in nature and do not make any assumptions concerning the generative distribution of the data. Instead, it is purely data driven. Adapting these techniques to a variety of data type requires defining a distance measure for the given data. With regards to mixed anomalies, semi-supervised techniques perform more improved than unsupervised techniques since the likelihood of an anomaly is to form a near neighbourhood when the training data set is low.

However, these techniques have disadvantages. They fail to label the anomalies correctly, resulting in missed anomalies, for unsupervised techniques. If the data has normal instances that do not have close neighbours or if the data has anomalies that have close neighbours the technique fails to label them correctly, resulting in missed anomalies. The computational

complexity of the testing phase is a challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data. In semi-supervised techniques, if the normal instances in the test data do not have enough similar normal instances in the training data, then the technique will have a high false positive rate.

6.2 Clustering-based anomalies detection techniques

Clustering-based anomalies detection techniques can be grouped into three assumptions:

The first assumption: *Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.* Techniques based on this assumption apply a known clustering-based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Several clustering algorithms do not force every data instance to belong to a cluster, such as *DBSCAN* (Ester *et al*., 1996), *ROCK* (Guha *et al.,* 2000) and *SNN clustering* (ErtÄoz *et a*l., 2003). The *FindOut* algorithm (Yu *et al*., 2002) is an extension of the *WaveCluster* algorithm (Sheik-holeslami *et al.,* 1998) in which the detected clusters are removed from the data and the residual instances are declared as anomalies. A disadvantage of these techniques is that they are not optimised to find anomalies, as the main aim of the underlying clustering algorithm is to find clusters. Typical application areas include image processing (Scarth *et a*l., 1995), and fraud detection (Wu and Zhang, 2003; Otey *et a*l. 2003).

The second assumption: *Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.* Techniques based on this assumption consist of two steps. In the first step, the data is clustered using a clustering algorithm. In the second step, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score. A number of anomaly detection techniques that follow this two-step approach have been proposed using different clustering algorithms. Smith *et al*. (2002) study *Self-Organizing Maps (SOM)*, *K-means* and *Expectation Maximization (EM)* to cluster training data and then use the clusters to classify test data. In particular, SOM (Kohonen, 1997) has been widely used to detect anomalies in a semi-supervised mode in several applications such as intrusion detection (Labib and Vemuri, 2002; Smith *et al*., 2002; Ramadas *et al*., 2003), fault detection (Harris, 1993; Ypma, Duin, 1998; Emamian *et al*., 2000) and fraud detection (Brockett *et al.,* 1998). Barbara *et al*. (2003) propose a robust technique to detect anomalies in the training data. This assumption can also operate in a semi-supervised mode, in which the training data are clustered, with instances belonging to the test data being compared against the clusters to obtain an anomaly score for the test data instance (Marchette, 1999; Wu and Zhang, 2003; Vinueza & Grudic, 2004; Allan *et al*., 1998). If the training data have instances belonging to multiple classes, semi-supervised clustering can be applied to improve the clusters to address this issue.

The third assumption: *Normal data instances belong to large and dense clusters, while anomalies belong either too small or too sparse clusters.* Techniques based on the above assumption declare instances belonging to cluster as anomalous if size/density is below a threshold. Several variations of the third assumption of techniques have been proposed (Pires and Santos-Pereira, 2005; Otey *et al*., 2003; Eskin *et al*., 2002; Mahoney *et al*., 2003; Jiang *et al*., 2001; He *et al*., 2003). The technique proposed by He *et al*. (2003), called *FindCBLOF*, assigns an anomaly score known as the Cluster-Based Local Outlier Factor (CBLOF) to each data instance. The CBLOF score captures the size of the cluster to which the data instance belongs, in addition to the distance of the data instance to its cluster centroid. These techniques are used for network intrusion detection (Bolton & Hand 1999), and for host based intrusion detection (Sequeira & Zaki 2002).

In terms of advantages these techniques can work in an unsupervised mode, and can be adapted to complex data types by working in a clustering algorithm that can handle the specific data type.

The testing stage for clustering based techniques is fast because the number of clusters against is a small constant. However these techniques are highly dependent on the effectiveness in capturing the cluster structure of normal instances. Numerous techniques detect anomalies as a result of clustering, and are not improved for anomaly detection. Some clustering algorithms are assigned to a particular cluster. This could result in anomalies getting assigned to a larger cluster, thus being considered as normal instances by techniques that work under the assumption that anomalies are not linked to any cluster. If O (N2d) clustering algorithms are used, then the computational complexity for clustering the data is often a bottleneck.

6.3 Statistical techniques

Statistical anomaly detection techniques are based on the following key assumption: ***Assumption:*** Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Statistical techniques operate in two phases: *training* and *testing* phases, once the probabilistic model is known. In the *training* phase, the first step comprises fitting a statistical model to the given data, whereas the *testing* phase, determines whether a given data instance is anomalous with respect to the model or not. This involves computing the probability of the test instance to be generated by the learnt model. Both parametric and non-parametric techniques are used. Parametric techniques assume the knowledge of underlying distribution and estimate the parameters from the given data (Eskin 2000). Non-parametric techniques do not assume any knowledge of distribution characteristics (Desforges *et al*., 1998). Typically the modelling techniques are robust to small amounts of anomalies in the data and hence can work in an unsupervised mode. Statistical techniques can operate in unsupervised settings, semi-supervised and supervised settings. Supervised techniques estimate the probability density for normal instances and outliers. The semi-supervised techniques estimate the probability density for either normal instances, or anomalies, depending on the availability of labels. Unsupervised techniques define a statistical model, which fits the majority of the observations. One such approach is to find the distance of the data instance from the estimated mean and declare any point above a threshold to be anomalies (Grubbs 1969). This requires a threshold parameter to determine the length of the tail, which has to be considered as anomalies; techniques used for mobile phone fraud detection (Cox *et al*., 1997).

The advantages of these techniques are as follows:

- If the assumptions concerning the underlying data distribution are true, these techniques then offer a statistically correct solution for anomaly detection.

- Confidence interval is associated with the anomaly score provided by a statistical technique, which can be used as extra information when making a decision concerning any test instance.

- It can operate in an unsupervised setting without any need for labelled training data if the distribution estimation step is robust to anomalies in data.

However, they rely on the assumption that the data is conducted from a particular distribution. This assumption is not necessarily true, particularly for high dimensional real data sets. Even when the statistical assumption can be justified, there are several hypothesis test statistics that can be useful to detect anomalies; choosing the greatest statistic is often not an easy task (Motulsky 1995). In specific, composing hypothesis tests for complex distributions needed to fit high dimensional data sets is nontrivial. An anomaly might have attribute values that are individually very common, but their combination is very uncommon, but an attribute-wise histogram based

technique would not be able to detect such anomalies. Histogram based techniques are relatively simple to apply, a key disadvantage of such techniques with regards to multivariate data is that they are not able to capture the interactions between different attributes.

6.4 Classification techniques

Classification based techniques operate under the following general assumption:

Assumption: A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space.

Classification is an important data-mining concept. The aim of classification is to learn a set of labelled data instances (training) and then classify an unseen instance into one of the learnt class (testing). Anomalies detection techniques based on classification also operate in the same two-phase, using normal and anomalies as the two classes. The training phase builds a classification model using the available labelled training data. The testing stage classifies a test instance using the model learnt. The techniques following this approach fall under supervised anomalies detection techniques. A one-class classifier can then be trained to reject this object and to label it as anomalies. These techniques fall under the category of semi-supervised anomalies detection techniques (Tan *et al*. 2005b; Duda *et al*. 2000).

The classification problem is modelled as a two-class problem where any new instance that does not belong to the learnt class is anomalous. In real scenarios, class labels for normal class are more readily available but there are also cases where only anomalies class labels are available. Classification based techniques are categorised into subcategories based on the type of classification model that use. These include Neural networks, Bayesian Networks, Support Vector Machines (SVM), decision trees and regression models. These rules are used to classify a new observation as normal or anomalous. In term of advantages, the testing stage of these techniques is fast since each test instance needs to be compared against the pre-computed model. They can make use of powerful algorithms that can differentiate between instances belonging to different classes. However, Multi-class classification techniques rely on availability of precise labels for different normal classes, which is often not possible. These techniques allocate a label to each test instance, which can become a disadvantage when a meaningful anomaly score is wanted for the test instances. Some classification techniques that obtain a probabilistic prediction score from the output of a classifier can be used to address this issue (Platt 2000).

6.5 Information Theory Based

These techniques are based on the following key assumption:

Assumption*:* Anomalies in data induce irregularities in the information content of the data set.
Information theory based techniques analyse the information content of a dataset using different information theoretic measures such as relative entropy, entropy, *etc*. The general idea is that normal data is regular in terms of a certain information theoretic measure. Anomalies significantly change the information content of the data because of their surprising nature. Thus, the typical approach adopted by this technique is to detect data instances that induce irregularity in the data, where the regularity is measured using a particular information theoretic measure. Information theory based techniques operate in an unsupervised mode.

The advantages of these techniques are as follows:

- They can function in an unsupervised setting.
- They make no assumptions regarding the underlying statistical distribution of the data.

However, the performance of these techniques is greatly dependent on the choice of the information theoretic measure. Frequently, these measures can detect anomalies only when there are large numbers of anomalies existing in the data. It is often nontrivial to obtain when these techniques are applied to sequences and spatial data sets because they rely on the size of the substructure. Another disadvantage is that it is difficult to associate an anomaly score with a test instance using these techniques.

6.6 Other Techniques

These techniques are based on the following key assumption:

Assumption: Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different.

Spectral decomposition based technique finds an approximation of the data using a combination of attributes that capture the size of variability in the data. The underlying assumption for such techniques is that the reduced sets of attributes faithfully capture much of the normal data, but this is not necessarily true for the anomalies. Spectral techniques can work in an unsupervised as well as semi-supervised setting. This approach has been applied to the network intrusion detection domain by several different groups (Shyu *et al*. 2003; Lakhina *et al.* 2005; Thottan and Ji 2003) and for detecting anomalies, for example in spacecraft components (Fujimaki *et al*. 2005).

Visualisation based technique maps the data in a coordinate space that makes it easy to visually identify the anomalies. Cox *et al.* (1997) present a visualisation-based technique to detect telecommunications fraud, which displays the call patterns of various users as a directed graph such that a user can visually identify abnormal activity.

These techniques routinely perform dimensionality reduction, which makes them suitable for handling high dimensional data sets. Additionally, they can be used as a pre-processing step, followed by application of any existing anomaly detection technique in the transformed space. These techniques can be used in an unsupervised setting.

However, these techniques usually have high computational complexity. They are useful only if normal and anomalous instances are separate in the lower dimensional embedding of the data.

6.7 Overview of strengths and limitations

For high-dimensional data, any of the above anomalies detection techniques can easily detect the anomalies. For more complex data sets, different techniques face different challenges. Chandola *et al*. (2009) argue that statistical techniques do not work well with high-dimensional categorical data and that visualisation-based techniques are more naturally suited to low-dimensional data and hence require dimensionality reduction as a pre-processing step when dealing with a higher number of dimensions. Spectral decomposition-based techniques, which find an approximation of the data using a combination of attributes to capture the variability in the data, explicitly address the high-dimensionality problem by mapping data to a lower dimensional projection, but their performance is highly dependent on the fact that the normal instances and anomalies are distinguishable in the projected space. Clustering is often called an unsupervised learning task, as no class values indicate an a priori grouping of the data instances, as in the case for supervised learning. Clustering and nearest neighbour techniques rely on a good similarity or distance measure to handle the anomalies in complex data sets. Classification-based techniques handle the dimensionality better, since they try to assign weights to each dimension and ignore unnecessary dimensions automatically. However, classification-based techniques require labels for both

normal data and anomalies. Finally, information theory-based techniques, which analyse the information content of a data set using different information theoretic measures (e.g. entropy measure), require a measure that is sensitive enough to detect the effects of even single anomalies. Such techniques detect anomalies only when there is a significant number of an anomaly.

7. CHALLENGES OF ANOMALIES DETECTION

Multi- and high-dimensional data make the outlier mining problem more complex because of the impact of the curse of dimensionality on algorithms' performance and effectiveness. Wei *et al*., (2003) introduce an anomalies mining method based on a hyper-graph model to detect anomalies in a categorical data set. He *et al*. (2005) define the problem of anomalies detection in categorical data as an optimisation problem from a global viewpoint, and present a local search heuristic-based algorithm for efficiently finding feasible solutions. He *et al*. (2005) also present a new method for detecting anomalies by discovering frequent patterns (or frequent item sets) within the data set. The anomalies are defined as the data transactions that contain less frequent patterns in their item sets. The recent surveys on the subject (Chandola *et al*., 2009; Patcha & Park, 2007) note that anomalies detection has traditionally dealt with record or transaction type data sets. They further indicate that most techniques require the entire test data before detecting anomalies, and mention very few online techniques. Indeed, most current algorithms assume that the data set fits in the main memory (Yankov *et al*., 2007). Both aspects violate the requirement for real-time monitoring data streams. In addition, most approaches focus specifically on intrusion detection (Kuang & Zulkernine, 2008; Xu *et al*., 2005; Lee & Stolfo, 2000). A comparative study (Chandola *et al.*, 2008) of methods for detecting anomalies in symbolic data shows that there are several techniques for obtaining a symbolic representation from a time series (Lin *et al*., 2007; Bhattacharryya & Borah, 2004), but all such works seem to apply solely to univariate data (Keogh *et al*., 2004; Wei *et al*., 2003). It is a challenging task to detect failures in large dynamic systems because anomalous events may appear rarely and do not have fixed signatures.

8. ANOMALIES DETECTION AND LINK MINING

The literature review reveals a growing range of applications in anomalies detection, mostly to data mining and very few applications in link mining. In recent years application of anomalies detection in link mining has gained increasing importance. For example, the paper of Savage *et a*l (2014) in online social networks survey's existing computational techniques used to detect irregular or illegal behaviour; other works include detecting fraudulent behaviour of online auctioneers (Chan *et al.,* 2006). Community based anomalies detection in evolutionary networks (Chen *et al*., 2012), link based approach for bibliometric journal ranking (Su *et al.,* 2013). However, their focus is still on pattern finding rather than link related tasks. Even the work on citation data (Keane, 2014, Yang *et al*., 2011) is used to describe communities or computational techniques and not mining anomalies or predictive links. Thus, much of the work in this area has focused on identifying patterns in behaviour of the data rather than link mining. Anomalies detection in link mining is still an emerging area.

9. SUMMARY

Link mining is an emerging area within knowledge discovery focused on mining task relationship by exploiting and explicitly modelling the links among the entities. We have overviewed link mining in terms of object related task, link-based object and group related task. These represent some of the common threads emerging from 9 varieties of fields that are exploring this exciting and rapidly expanding field. However, with the introduction of links, new tasks also come to light: predicting the type of link between two objects, predicting the numbers of links, inferring

the existence of a link, and inferring the identity of an object. A review of computational techniques is provided outlining their challenges. Anomaly detection, which is important to use in this research, is also discussed and the current methods and issues highlighted. These two areas are attracting much interest by researchers from different disciplines (*e.g.* computer science, business, statistics, forensics and social sciences) interested in extracting tacit, hidden, but valuable knowledge from the vast amount of data available worldwide.

REFERENCES

[1] Abe S., Kawano H., Goldstein J., Ohtani S., Solovyev S.I., Baishev D.G. and Yumoto K. (2006) Simultaneous identification of a plasmaspheric plume by a ground magnetometer pair and IMAGE Extreme Ultraviolet Imager. Journal of Geophysical Research 111(A11).

[2] Aggarwal C., and Yu P.(2001) Outlier Detection for High Dimensional Data. International Conference on Management of Data. 30(2). P.37 – 46.

[3] Aggarwal R, Isil E, Miguel A. Ferreira, and Matos P.(2011) Does Governance Travel Around the World? Evidence from Institutional Investors, Journal of Financial Economics 100. P.154-181.

[4] Aggarwal R., Gehrke J., Gunopulos D.,and Raghavan P.(1998) Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proceedings of the ACM SIGMOD International Conference on Management of Data. 27(2). p.94 – 105.

[5] Aggarwal Y. Zhao, and Yu P.S.(2011) Outlier Detection in Graph Streams, ICDE Conference.

[6] Allan J., Carbonell J., Doddington G., Yamron J., and Yang Y. (1998) Topic detection and tracking pilot study. Proceedings of DARPA Broadcast News Transcription and Understanding Workshop.

[7] Badia A., Kantardzic M.(2005) Link Analysis Tools for Intelligence and Counterterrorism. ISI. P.49-59.

[8] Barbara D., Li Y., Couto J., Lin J. L., and Jajodia S.(2003) Bootstrapping a data mining intrusion detection system. Proceedings of the 2003 ACM symposium on Applied computing. ACM Press.

[9] Bhattacharyya N, Bandyopadhyay R, Bhuyan M, et al (2005) correlation of multi-sensor array data with taster's panel evaluation. Proceedings of ISOEN, Barcelona, Spain.

[10] Brockett P. L., Xia, X., and Derrig R. A.(1998) Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance. 65(2) P.245-274.

[11] Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection. A Survey, ACM Computing Survey. 41(3). p.15.

[12] Chandola V., Eilertson E., Ertoz L., Simon,G., and Kumar V.(2006) Data mining for cyber security. Data Warehousing and Data Mining Techniques for Computer Security, A. Singhal, Ed. Springer.

[13] Chau D. H., Pandit S., Faloutsos C.(2006) Detecting fraudulent 1032 personalities in networks of online auctioneers. In: Knowledge Discovery in Databases: PKDD.p.103–114.

[14] Chellappa., Rama J., and Anil.(1993) Boston: Academic Press.

[15] Chen Z., Hendrix W., Samatova N. F.(2012) Community-based anomaly detection in evolutionary networks. Journal of Intelligent Information Systems. 39(1).p.59–85.

[16] Cox C., Enno A., Deveridge S., Seldon M., Richards R., Martens V., and Woodford P.(1997) Remote electronic blood release system. Transfusion.37.p.960-974.

[17] Creamer, G., and Stolfo, S.(2009) A link mining algorithm for earnings forecast and trading Data. Min Knowl Disc. 18. P.419–445.

[18] Desforges D. M., Lord C. G., Ramsey S. L.(1998) Effects of structured cooperative contact on changing negative attitudes toward stigmatized social groups. Journal of Personality and Social Psychology.60.p.531 -544.

[19] DesJardins M., and Matthew E.(2006) Gaston, Speaking of relations: Connecting statistical relational learning and multi-agent systems. ICML Workshop on Open Problems in Statistical Relational Learning, Pittsburgh, PA.

[20] Domingos P., Doan AH., Madhavan J., and Halevy A.(2004) Ontology matching: A machine learning approach. Handbook on ontologies.

[21] Duda R. O., Hart P. E., and. Stork D. G. (2000) Pattern Classification and Scene Analysis, John Wiley & sons.

[22] Emamian V., Kaveh M., and Tewfik A.(2000) Robust clustering of acoustic emission signals using the kohonen network. Proceedings of the IEEE International Conference of Acoustics,Speech and Signal Processing. IEEE Computer Society.

[23] ErtÄoz A., Arnold M., Prerau L., Portnoy., and Stolfo S.(2003) A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Proceedings of the Data Mining for Security Applications Workshop.
[24] Ertoz L.; Steinbach, M.; Kumar V.(2004). Finding Topics in collections of documents: A shared nearest neighbour approach. Clustering and Information Retrieval. P.83-104.
[25] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S.(2002) A geometric framework for unsupervised anomaly detection. Proceedings of Applications of Data Mining in Computer Security. Kluwer Academics.P.78-100.
[26] Ester M., Kriegel H-P., Sander J., and Xu X.(1996) A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. P.226 – 231.
[27] Fawcett T., Provost F.(1999) Activity monitoring: noticing interesting changes in behavior. Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99).p.53–62.
[28] Fujimaki R., Yairi T., and Machida K.(2005) An approach to spacecraft anomaly detection problem using kernel feature space. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York.p.401–410.
[29] Getoor L.(2003). Link mining. A new data mining challenge, SIGKDD Explorations, 5(1). p.84-89.
[30] Getoor L.(2005) .Tutorial on Statistical Relational Learning. ILP: 415.
[31] Getoor L., and Diehl C.(2005). Link mining: A survey SIGKDD Explorations, December. Vol.7 (2).
[32] Ghosh S., and Reilly D. L.(1994). Credit card fraud detection with a neural-network. Proceeding of the 27th Annual Hawaii International Conference on System Science.3.
[33] Goldberger, a.l.,amaral,A.N.,Glass,L.,Havs dorff,J.M.,ivanov,pc.,mark,R.G.,et al.(2000) physiobank, physiotoolkit and phyionet.circulation,101, 215-220.
[34] Grubbs Frank E.(1969) Procedures the data to assure that the results for detecting outlying observations in are representative of the thing samples. Technometrics 11.p.1-2.
[35] Guha S., Rastogi R., and Shim K.(2001). ROCK: A robust clustering algorithm for categorical attributes. Information Systems 25(5). p. 345-366.
[36] Han J., and Kamber M.(2001) Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers. P.550.
[37] Harris T. (1993). Neural network in machine health monitoring. Professional Engineering.
[38] Harvey Motulsky (1995). Intuitive Biostatistics. Newyork: Oxford University Press. 200-386.
[39] Hu T., and Sung S.Y.(2003) Detecting pattern-based outliers. Pattern Recognition Letters.24 (16). P.3059 – 3068.
[40] Jin W., Tung A., and Han J.(2001). Mining Top-n Local Outliers in Large Databases. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.p.293 – 298.
[41] Keane M., (2014). (Big) Data Analytics: From Word Counts to Population Opinions. insight. 1,1-45.
[42] Keogh E., Lonardi S., and Ratanamahatana C. A.(2004). Towards parameter-free data mining. Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press.p. 206-215.
[43] Kirkland, D., Senator, T., Hayden, J., Dybala, T., Goldberg, H. & Shyr, P. (1999). The NASD Regulation Advanced Detection System. AAAI 20(1): Spring, 55-67.
[44] Kohonen T.(1997) Self-organizing maps. Springer-Verlag New York, Inc.
[45] Kuang L., and Zulkernine M.(2008) An anomaly intrusion detection method using the CSI-KNN algorithm. SAC.P. 921-926.
[46] Labib K., and RaoVemuri V.(2002) “NSOM: A Real-time Network-Based Intrusion detection System Using Self-Organizing Maps, Networks and Security.
[47] Lakhina A., Crovella M., and Diot C.(2005) Mining Anomalies Using Traffic Feature Distributions. Proceedings of ACM SIGCOM.p. 217-228.
[48] Lee W., and Stolfo, S.(2000) A framework for constructing features and 638 models for intrusion detection systems. ACM Transactions on 639 Information and System Security. 3(4).
[49] Liben-Nowell D., and Kleinberg J.(2003) The link prediction problem for social networks. In CIKM ’03. Proceedings of the twelfth international conference on Information and knowledge management.p.556–559.
[50] Lin H., Fan W., and Wallace L.(2007) An empirical study of web-based knowledge community success. Proceedings of the 40th Hawaii International Conference on System Sciences. P.1530-160.

[51] Lin S., and Brown D.(2004) An Outlier-based Data Association Method for Linking Criminal Incidents. Proceedings of the SIAM International Conference on Data Mining.

[52] Lin S., and Brown D.(2003) An Outlier-based Data Association Method. Proceedings of the SIAM International Conference on Data Mining.

[53] Marchette D.(1999) A statistical method for profling network traffic. Proceedings of 1st USENIX Workshop on Intrusion Detection and Network Monitoring. Santa Clara, CA, .p.119-128.

[54] Marcus G., Fernandes K., and Johnson S.(2007) Infant rule-learning facilitated by speech. Psychol. Sci.18.p.387–391.

[55] Mustafa Y. T., Tolpekin V., and Stein A., and Sub M.(2007) The application of Expectation Maximization algorithm to estimate missing values in Gaussian Bayesian network modeling for forest growth. IEEE Transactions on Geoscience and Remote Sensing.

[56] O'Madadhain J., Smyth P., and Adamic L.(2005) Learning Predictive Models for Link Formation. To be presented at the International Sunbelt Social Network Conference.

[57] Otey M. E., Ghoting A., and Parthasarathy S.(2003) Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery. 12(2-3) p.203-228.

[58] Otey M., Parthasarathy S., Ghoting A., Li G., Narravula S., and Panda D.(2003).

[59] P´oczos Z., and L˝orincz A.(2009) Complex independent process analysis. PLA University of Science & Technology, Nanjing 210007, China.

[60] Panzeri S., Brunel N., Logothetis NK., and Kayser C.(2010) Sensory neural codes using multiplexed temporal scales. Trends Neurosci.33.p.111–120.

[61] Patcha A., and Park JM.(2007) An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks. 51(12).p.3448-3470.

[62] Petrovskiy M.(2003) Outlier Detection Algorithms in Data Mining Systems. Programming and Computing Software. 29(4).p.228 – 237.

[63] Pires A., and Santos-Pereira C.(2005) Using clustering and robust estimators to detect outliers in multivariate data. Proceedings of International Conference on Robust Statistics. Finland.

[64] Platt J.(2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds.p.61–74.

[65] Popescul A., Ungar L., Lawrence S., and Pennock D.(2003) Statistical re-lational learning for document mining. Computer and Information Sciences, University of Pennsylvania.

[66] Provana K.G., Leischowc S. J., Keagyb J., and Nodorac J.(2010) Research collaboration in the discovery, development, and delivery networks. of a statewide cancer coalition.33(4).p. 349-355.

[67] Ramadas M., Ostermann S., and Tjaden B. C.(2003) Detecting anomalous network traffic with self-organizing maps. Proceedings of Recent Advances in Intrusion Detection.P.36-54.

[68] Ramaswamy S., Rastogi R., and Shim K.(2000). Efficient Algorithms for Mining Outliers from Large Data Sets. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 29(2).p.427 – 438.

[69] Rattigan M. J., and Jensen D.(2005) The case for anomalous link discovery. SIGKDD Explorations, 7(2).

[70] Savage D., Zhang X., Yu X., Chou P., and Wang Q.(2014) Anomaly Detection in Online Social Networks. Social Networks.39.p.62–70.

[71] Scarth G., McIntyre M., Wowk B., and Somorjai R.(1995) Detection of novelty in functional images using fuzzy clustering. Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine. Nice, France.p.238.

[72] Sequeira K. and Zaki M.(2002) Admit: anomaly-based data mining for intrusions. In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press.p.386–395.

[73] Sheikholeslami G., Chatterjee S., and Zhang A.(1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc.p.428-439.

[74] Shyu M.L., Chen S.C., Sarinnapakorn K., and Chang L.(2003) A novel anomaly detection scheme based on principal component classifier. Proceedings of 3rd IEEE International Conference on Data Mining.p.353–365.

[75] Skillicorn D. B.(2004) Detecting Related Message Traffic, Workshop on Link Analysis, Count ErtÄoz errorism, and Privacy. SIAM International Conference on Data Mining, Seattle, USA.

[76] Smith R., Bivens A., Embrechts M., Palagiri C., and Szymanski B.(2002) Clustering approaches for anomaly based intrusion detection. Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. ASME Press.P.579-584.

[77] Song X., Wu M., Jermaine C., and Ranka S.(2007) Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering.19(5).p.631-645.
[78] Sparrow M.(1991) The application of network analysis to criminal intelligence: an assessment of the prospects. Soc Netw 13.p.251–274.
[79] Su, P., Shang C., and Shen A.(2013) Soft Computing - A Fusion of Foundations, Methodologies and Applications archive. 17(12).p.2399-2410.
[80] Su, P., Shang C., and Shen A.(2013) "Link-based approach for bibliometric journal ranking," Soft Computing, to appear.
[81] Tan L.,Taniar D., and Smith K.(2005) Introduction to Data Mining. Addison-Wesley.J(2).p.229-245.
[82] Taskar B., Abbeel P., and Koller D.(2003) Discriminative probabilistic models for relational data. Proc. UAI02, Edmonton, Canada.
[83] Thottan., and Ji.(2003) Anomaly detection in IP networks. Signal Processing, IEEE Transactions .51(8).p.2191-2204.
[84] Vinueza A., and Grudic G.(2004) Unsupervised outlier detection and semi-supervised learning.Tech. Rep. CU-CS-976-04, Univ. of Colorado at Boulder.
[85] Wadhah,A.,Gao,S., Jarada,T., Elsheikh,A.,Murshed,A.,Jida,J.,Alhajj,R.. (2012). Link prediction and classification in social networksand its application in healthcare and systems biology. Netw Model Anal Health Inform Bioinforma. 1 (2), 27-36.
[86] Wei L., Qian W., Zhou A., and Jin W.(2003). Hot: Hypergraph-based outlier test for categori-cal data. Proceedings of the 7th Pacic-Asia Conference on Knowledge and Data Discovery. p.399-410.
[87] Wu N. , and Zhang J.(2003) Factor analysis based anomaly detection. Proceedings of IEEE Workshop on Information Assurance. United States Military Academy, West Point, NY.
[88] Wu J., Xiong H., and Chen J.(2009) .Adapting the right measures for k-means clustering, in KDD.p.877–886.
[89] Xu, K.M., Zhang M, Eitzen Z.A., Ghan S.J., Klein S.A., and Zhang J.(2005) Modeling springtime shallow frontal clouds with cloud-resolving and single-column models. J. Geophys. Res., 110, D15S04, doi:10.1029/2004JD005153.
[90] Yankov D., Keogh E. J., and Rebbapragada U. (2007). Disk aware discord discovery: Finding unusual time series in terabyte sized datasets. Proceedings of International Conference on Data Mining.p.381-390.
[91] Yang Y., Zhiguo G., and Leong H.U.(2011). Identifying points of interest by self-tuning clustering. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11). ACM, New York.
[92] Ypma A., and Duin R.(1998). Novelty detection using self-organizing maps. Progress in Connectionist Based Information Systems.2.p.1322-1325.
[93] Yu D., Sheikholeslami G., and Zhang A.(2002) Findout: finding outliers in very large datasets. Knowledge And Information Systems.4(4).p. 387.

INTENTIONAL BLANK

MUSIC INCORPORATING UTERINE CONTRACTION IN NON-INVASIVE FETAL HEARTBEAT DETECTION

Walid A. Zgallai

Faculty of Engineering Technology and Science,
Higher Colleges of Technology, UAE
walidzgallai@yahoo.co.uk

ABSTRACT

The aim of this paper is to detect fetal heart beats temporally overlapping with the transabdominally-measured QRS-complexes of the mother, non-invasively. Modified, Weighted and uterine contraction interference signal covariance matrix incorporated spectral MUSIC technique is applied. It is based on partitioning the subspace containing the ECG signal bearing the mother and fetal, and the orthogonal subspace containing the uterine contraction interference signal plus noise. This exploits the orthogonality between the signal and noise subspaces provided that the noise is additive white Gaussian. In the modified MUSIC, subsequent separation of the mother and fetal QRS-complexes is performed in their shared signal subspace.

KEYWORDS

Non-invasive, Fetal heartbeat detection, MUSIC Spectral estimation, uterine contraction, Weighted Kaiser filter.

1. INTRODUCTION

Difficult situations arise in which the maternal and fetal heart beats are commensurate. Episodes of near coincident maternal and fetal QRS-complexes have been found in about 10% of the transabdominal ECG data. In such episodes about every ten seconds a fetal heartbeat coincides with the maternal QRS-complex. This is similar to one problem which has often arisen in Radar applications [1] where two coincident targets have common temporal and spectral characteristics. Such a problem and others different in nature, e.g., Sonar, and underground buried objects, have been dealt with using the following spectral estimation methods that are based on partitioning the signal and noise subspaces; (i) the conventional multiple signal classification (MUSIC) method [2], (ii) the Pisarenko harmonic decomposition (PHD) method [3], (iii) the eigenvector (EV) method [4], and (iv) the minimum norm method [5]. Such subspace parameter or frequency estimation methods differ only in what part of the noise subspace they each use [6].

The maternal QRS-complex principal spectral peak is around 17 Hz, and the fetal QRS-complex principal spectral peak is around 30 Hz [7]. The spectral content can be used in the detection of either signal within the maternal cardiac cycle. A modified MUSIC algorithm has been devoted to identifying anomalous QRS-complexes and P-waves such as P-on-T-waves and P-on-QRS-complex episodes for adult patients in the frequency domain [7]. For fetal heart rate (FHR) detection in labour one has to overcome two problems; (i) poor spectral resolution, and (ii) the

David C. Wyld et al. (Eds) : CCSEA, CLOUD, DKMP, SEA, SIPRO - 2016
pp. 191–198, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60217

influence of the coexisting labour contraction signals [8] which exhibits a broad spectrum, and are characterised by having resonances, one of which is overlapping with the main fetal spike event. The fetal heartbeat detection is accomplished by thresholding the enhanced fetal spikes in the frequency domain. A challenge is to enhance the resolution of the mother and fetal QRS-complexes' principal pseudo-spectral peaks, (MPPP) and (FPPP), respectively, and to nudge the UCS plus noise into a separate subspace, the interference subspace (I-subspace), whereby orthogonalisation is forced between the I-subspace and the signal subspace (S-subspace) containing both the mother and / or the fetal QRS signature imprints. An auxiliary method based on *the* Gram-Schmidt orthogonalisation is employed in addition to Generalised Singular Value Decomposition (GSVD) which deals with partitioning signal and coloured noise subspaces [9]. This technique deals with the UCS during the strong peaks of labour contractions which have noise-like characteristics and are heavily contaminated with other noise artefact. The paper is organised as follows; Methodology is described in section 2. Results are shown in Section 3. Discussion is presented in Section 4. Conclusions are summarized in section V.

2. METHODOLOGY

Each maternal cardiac cycle has been divided into four segments of 250 msec each. The segmentation starts 50 msec before the maternal R-wave and continues until the end of the first segment. The other three equal segments are adjusted according to the maternal heart rate. There are inevitable deviations in the 17 Hz and the 30 Hz of the mother and fetal QRS-complex pseudo-spectra, respectively. Five overlapping and optimised Kaiser windows have been used in the detection of the MPPPs; 15-19 Hz. Ten overlapping and optimised Kaiser weighted windows have been used in the detection of the FPPPs; 28-38 Hz. The optimised Kaiser weights have been given in [10]. The model order has to be chosen carefully. The optimum model order is found by trial and error to be eleven for the signal and four for the noise. The method is not sensitive to small deviations in the model order.

The spectrum of the UCS may include comparatively strong narrowband spectral components centred around 5 Hz, 30 Hz, 45 Hz, 60 Hz, and 90 Hz in addition to some broadband components [10]. The uterine contraction component at 30 Hz masks the FPPP. A challenge is isolating the FPPP at 30 Hz in the presence of the UCS peak at the same frequency. Using a new pseudo-spectral localiser which incorporates the modified covariance matrix representing the UCS plus coexisting noise artefact, and seeks to reduce the influence of background uterine activities in the pseudo-spectral MUSIC localisation procedure by partitioning the two subspaces; one contains the desired signal parameters and the other contains the UCS parameters, is proposed. An accurate estimate of the UCS modified covariance matrix is needed to be incorporated in the pseudo-spectral localiser. A portion of the data that contains only noise fields, and does not contain any signal information such as the P-waves or the QRS-complexes, is utilised. When such a segment of the data, that is P-wave- and QRS-complex-free, is sufficiently long for the MUSIC pseudo-spectral localiser, an accurate estimate of the UCS modified covariance matrix can be obtained.

The mathematical formulation is based on [4, 10]. A flowchart is given in [10]. The temporal window is restricted to 250 msec. The Kaiser filter weights are applied to each of the 250 msec windows and the weights are optimised to enhance the principal peaks of either QRS-complex in their respective temporal domains. The data portions earmarked for the I_{noise} are segments that are free from mother and fetal QRS-complexes.

To exploit a MUSIC methodology [5] which incorporates a tailor-made subspace fitting for individual QRS spectral signatures based on *a priori* information, if we ignore the influence of the uterine contraction interference signals, the technique is based on weighting the covariance

matrix of the transabdominally-measured signals, which in turn uniquely modifies the signal and noise subspaces to enhance and retain only eigenvectors that result in the MPPP at 17 Hz, or the FPPP at 30 Hz. In the absence of uterine contraction interference signals and assuming white Gaussian noise presence, this is a weighted MUSIC technique. The signal and noise subspaces will be reconfigured by two tailor-made weighting Kaiser functions, one is aimed at enhancing the maternal QRS spectral peak and the other is aimed at enhancing the fetal QRS spectral peak.

3. RESULTS

The proposed localiser is applied to segments of the trnasabdominally-measured maternal ECG signal. Linearisation of the data is employed [11]. The UCS modified covariance matrix is calculated using the data portion in the segments that are maternal and fetal event free. Fig. 1 shows the results using the sequentially optimised, weighted MUSIC with and without the incorporation of the UCS modified covariance matrix for the case of maternal and fetal R-wave separation of 9 msec. Fig. 1 (a) depicts superimposed and synchronised maternal transabdominal and fetal scalp ECGs with maternal R-wave to fetal R-wave separation of 9 msec, respectively. Fig. 1 (b) shows the results employing the Modified MUSIC without UCS incorporation. The maternal MPPP is at 17 Hz, shown at the left hand part of the figure, the FPPP of the first fetal heartbeat is shifted at 31 Hz, shown in the inset at the right hand part of the figure. Fig. 1 (c) depicts the results of the Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC for the transabdominally-measured ECG signal. The fetal FPP is stronger and sharper around 31 Hz, and there is significant noise reduction in the QRS-free segments [7].

The effect of proximity of the maternal and fetal R-waves on the frequency deviation of the FPPP around 30 Hz, and on the fetal heart detection rate, in all observed cases of coincident mother and fetal QRS-complexes has been studied. The proposed algorithm has been applied to approximately 50,000 maternal cardiac cycles, including 4,873 coincident QRS-complexes cases. The results are shown in Table 1.

TABLE 1: The effect of proximity of the maternal and fetal R-wave on the frequency deviation of the FPPP at 30 Hz, and on the fetal detection rate.

R_m-R_f separation	40 msec	35 msec	25 msec	20 msec	15 msec	7 msec	0 msec
Frequency deviation ±	1.73 Hz	1.92 Hz	2.09 Hz	2.17 Hz	2.31 Hz	2.52 Hz	2.74 Hz
Overlapping windows	5	5	5	5	8	9	10
Detection rate (%)	93.81	93.63	93.56	93.49	93.24	92.35	91.83

From the overall results, it is observed that;

1. For a fixed model order of 11 and 4 for the signal and noise subspaces, respectively, the algorithm is capable of detecting fetal heartbeats, at a rate of 92%, when the mother and fetal R-waves are synchronised, provided that appropriate sequential weightings for the mother and the fetal are maintained throughout. As the separation between the mother and fetal R-waves is increased, there is a slight increase in the corresponding detection rate and a decrease in the FPPP frequency deviations.

2. The incorporation of the covariance matrix of the UCS helps to strengthen and sharpen the FPPPs in some cases and hence improves the resolution, and reduces the sensitivity of the FPPPs to small deviations from the optimal model order.

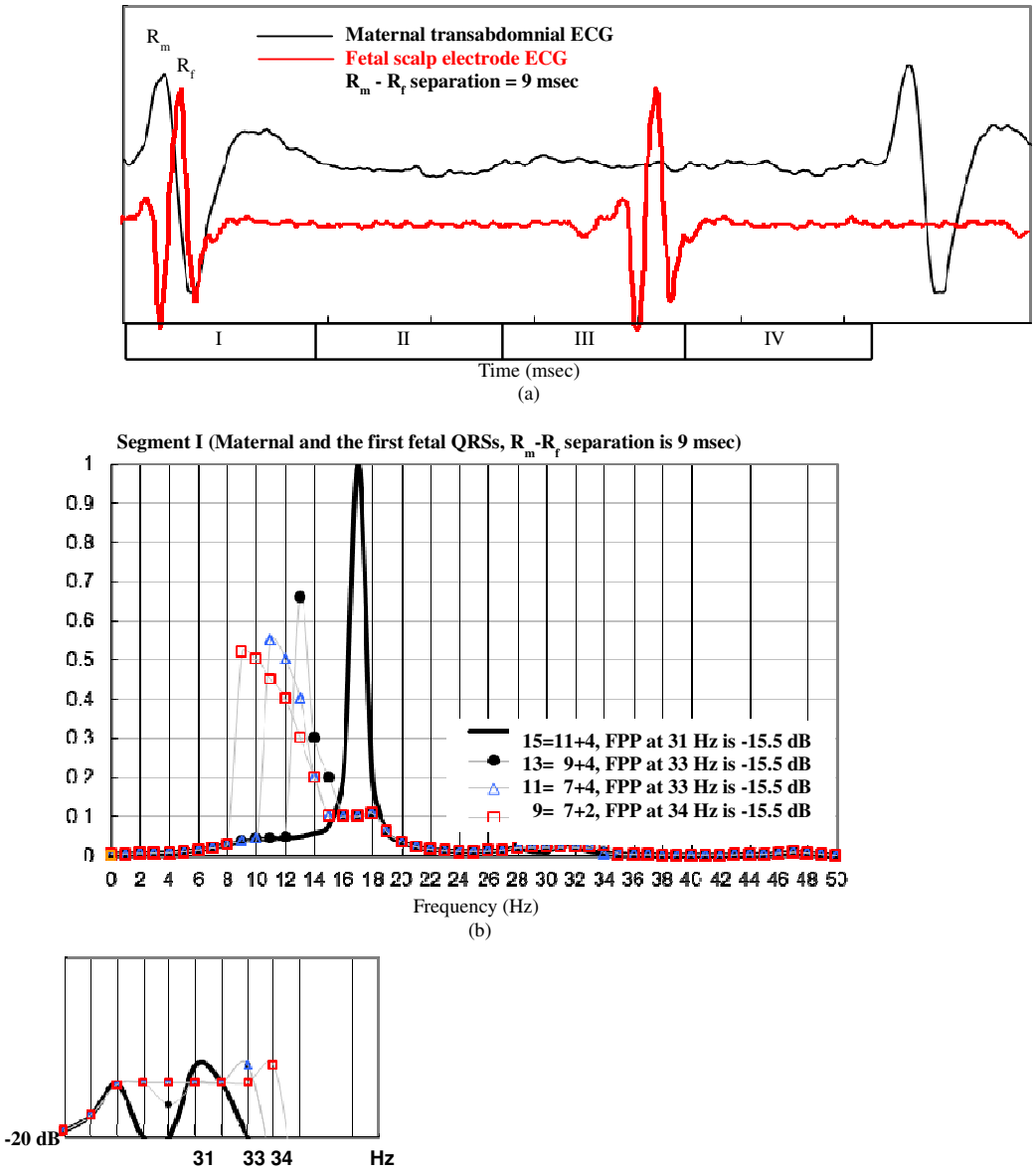
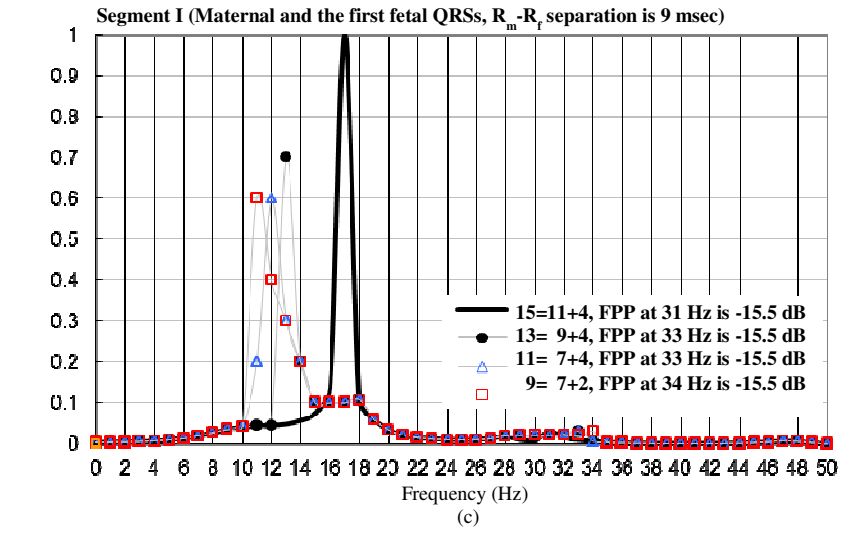


Figure 1. (a) Superimposed and synchronised maternal transabdominal and fetal scalp ECGs with maternal R-wave to fetal R-wave separation of 9 msec. The maternal cardiac cycle begins 50 msec before the R-wave and ends 50 msec before the next R-wave. The subject is at the first stage of labour, 40 weeks gestation. The maternal cycle has 500 samples at a sampling rate of 0.5 KHz. Segment I: maternal QRS-complex, segment II: the first fetal heartbeat with maternal contribution, segment III: QRS-free ECG, and segment IV: the second fetal heartbeat with maternal contribution. (b) Weighted spectral MUSIC for segment I of the transabdominally-measured ECG signal. As a result of close proximity, the FPP tends to broaden. Also, the FPP exhibits increased sensitivity to small deviations from the optimal model order in segment I. (c) Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC of segment I. Insets (right) show the FPPPs in dB.



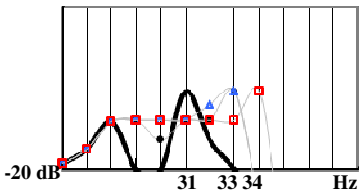


Figure 1. (continued) (a) Superimposed and synchronised maternal transabdominal and fetal scalp ECGs with maternal R-wave to fetal R-wave separation of 9 msec. The maternal cardiac cycle begins 50 msec before the R-wave and ends 50 msec before the next R-wave. The subject is at the first stage of labour, 40 weeks gestation. The maternal cycle has 500 samples at a sampling rate of 0.5 KHz. Segment I: maternal QRS-complex, segment II: the first fetal heartbeat with maternal contribution, segment III: QRS-free ECG, and segment IV: the second fetal heartbeat with maternal contribution. (b) Weighted spectral MUSIC for segment I of the transabdominally-measured ECG signal. As a result of close proximity, the FPP tends to broaden. Also, the FPP exhibits increased sensitivity to small deviations from the optimal model order in segment I. (c) Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC of segment I. Insets (right) show the FPPPs in dB.

3. The modified MUSIC without UCS incorporation has resulted in the following fetal heart detection rates: (i) 89.23%, 97.51%, and 91.20% for coincident, non-coincident mother and fetal QRS-complexes, and overall average, respectively. The Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC has resulted in the following fetal heart detection rates: (i) 93.52%, 99.35%, and 95.50% for coincident, non-coincident mother and fetal QRS-complexes, and overall average, respectively. The results have been verified by the recording of the instantaneous scalp fetal heart rate, measured when deemed necessary by the doctor on call after consent is obtained.

To calculate the bias, the expected values of the estimates are those obtained using the 250 msec segments from the maternal transabdominal ECG signal for a predominantly maternal QRS segment and a fetal heartbeat with maternal contribution. Those true values and estimates were calculated for 1000 segments. The results are 1.23 and 2.15 for MPPPs and FPPPs, respectively. For the maternal and fetal QRS-complex, the more deviation of the detected frequency of the MPPP around 17 Hz and the FPPP around 30 Hz, respectively, from the respective actual frequency, the higher the bias will be. The variance range is 0–8, with an average of 4.127, when calculated for 120,000 FHBs.

4. DISCUSSION

Assuming a maternal heart rate of 60 bpm yields a cardiac cycle length of 1000 msec. Each maternal cardiac cycle has been divided into four equal segments of 250 msec. The average rate by which the first fetal event coincides with the QRS-complex of the mother is 9.8%, based on 50,000 maternal cardiac cycles. When the two QRS-complexes of the mother and fetal coincide in segment I, segment II is usually free from such events and may be taken as the UCS plus noise artefact segment. On average, the second fetal heartbeat occurs in segment III. And if there is a third fetal heartbeat, then it is likely to occur over both the fourth segment of the present cycle and the first segment of the next cycle. In most cases, two fetal heartbeat occurrences within each maternal cardiac cycle were encountered, even when the maternal heart rate goes up during labour contractions. The deceleration of the fetal heart rate after the peak of labour contractions is normal and not proven to be related to the maternal heartbeat as her heart will still be racing for a while after the peak of contractions.

Successful detection of coincident mother and fetal QRS-complexes has resulted in an increase of 9.3% and 5.4% over and above the cumulants [12] and the bispectrum [13] template matching techniques, respectively. The mother and fetal QRS-complexes coincide making it difficult to separate them using any time-domain technique. With the cumulants method [12] there is a 13.8% failure rate, partially due to 9.8% rate of QRS-complex coincidences, and the rest, 4% rate, is due to overlapping fetal QRS-complex and maternal T-wave. The bispectrum method [13] failure rate of 9.8% is purely due to QRS-complex coincidences as there is a shortcoming in acquiring sufficiently high resolution to separate the bispectral peaks of the mother and fetal QRS-complexes. The overlapping of the fetal QRSs and the maternal T-waves can be resolved by the bispectrum template matching technique. The above percentages of QRS-complex coincident episodes have been found in the 50,000 maternal heartbeat database. The alternative is to try to resolve them in the frequency-domain [14-15].

5. CONCLUSIONS

This paper proposed a modified, Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC technique to detect temporally overlapping fetal heartbeats with maternal QRS complexes from transabdominal measurements, with the fetal scalp as a reference when deemed necessary, during temporally and spectrally overlapping uterine contractions. Performance analysis showed increased rate of detection for the Weighted and $\mathbf{I}_{noise}$ incorporated spectral MUSIC over and above that achieved employing the weighted spectral MUSIC.

The incorporation of the covariance matrix of the UCS helps to strengthen and sharpen the FPPPs for the optimum model order and in some cases it appears to be tolerant to a change in the model order from 11 and 4 to 9 and 4 for the signal and noise subspace, respectively. It has also resulted in a significant noise artefact reduction in the QRS-free segments. The method has resulted in the following fetal heart detection rates: (i) 93.52% for coincident mother and fetal QRS-complexes, (ii) 99.35% for non-coincident mother and fetal QRS-complexes, and (iii) 95.50% overall average. Without the incorporation of the UCS modified covariance matrix into the mathematical formulation of the sequentially optimised, weighted MUSIC, the following fetal heart detection rates have been obtained: (i) 89.23% as opposed to the 93.52% for coincident mother and fetal QRS-complexes, (ii) 97.51% as opposed to the 99.35% for non-coincident mother and fetal QRS-complexes, because in the former no appropriate noise model was assumed in the analysis, and (iii) 91.20% overall average.

REFERENCES

[1] S. Haykin, A. Steinhardt, "Adaptive Radar detection and estimation," A volume in the WileySeries in Remote Sensing, J. A. Kong, Series Editor, J. Wiley and Sons, Inc., 1992.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," Proceedings RADC Spectrum Estimation Workshop, pp. 243-258, 1979.

[3] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," Geophysics Journal Royal Astronomy Society, Vol. 33, pp. 347-366, 1973.

[4] S. Haykin, Adaptive filter theory, Prentice Hall, 1991.

[5] S. Marple, Spectral Analysis With Applications, Prentice Hall, 1987.

[6] S. Kay, Modern Spectral Estimation: Theory and Applications, Prentice Hall, 1987.

[7] M.S.Rizk, et al, "Novel decision strategy for P-wave detection utilising nonlinearly synthesised ECG components and their enhanced pseudospectral resonances," IEE Proceedings Science, Measurement and Technology, Special section on Medical Signal Processing, vol. 147, No. 6, pp. 389-397, November 2000.

[8] M.S.Rizk, et al. Non-linear dynamic tools for characterising abdominal electromyographic signals before and during labour. Transaction on Instrumentation, Measurement and Control, vol. 22, pp. 243-270, 2000.

[9] J.C.Principe, D. Xu, and C. Wang, "Generalised Oja's rule for linear discriminant analysis with Fisher criterion," Proceedings of the International Conference Acoustics, Speech, & Signal Processing, pp. 3401, 3404, 1997.

[10] W.A.Zgallai, "Advanced Robust Non-Invasive Fetal Heart Detection Techniques During Active Labour Using One Pair of Transabdominal Electrodes", PhD Thesis, City University, UK, 2007.

[11] W.A.Zgallai, "The application of adaptive LMF quadratic and cubic Volterra filters to ECG signals," International Journal of Computer Theory & Engineering, Badawy, W. Ed., IACSIT Press. Vol. 7, No. 5, pp. 337-343, October 2015.

[12] W.A.Zgallai, Second- and Third-Order Statistical Characterization of Non-Linearity and Non-Gaussianity of Adult and Fetal ECG Signals and Noise, Chapter 2 in Practical Applications in Biomedical Engineering, Andrade, et al., Eds., ISBN 9789535109242,. January 9, 2013.

[13] W.A.Zgallai, Detection and Classification of Adult and Fetal ECG Using Recurrent Neural Networks, Embedded Volterra and Higher-Order Statistics, Chapter 11 in Recurrent Neural Networks and Soft Computing, ElHefnawi, Ed., InTech Open, ISBN 9799533075463. 2012.

[14] M.Rizk, et al. "Modified MUSIC Pseudospectral Analysis Reveals Common Uterus and Fetal Heart Resonances During Labour Contractions", the 22nd IEEE EMBS, EMB2000, USA, 23-28/7/2000.

[15] W.A. Zgallai, "MUSIC fetal heartbeat detection during uterine contraction," International Conference on Biomedical Engineering and Environmental Technology, London, UK, 21-22/3/2015.

AUTHORS

Dr. Walid A. Zgallai

Dr. Zgallai was born in Tripoli in 1969. He received his BEng in Electrical and Electronic Engineering from Kuwait University, Kuwait, Kuwait, 1992, BSc Degree in Electrical Engineering from Tripoli University, Tripoli, Libya in 1992. He received his MSc degree in Communications and Signal Processing from the University of London, London, UK in 1993 and a Diploma of Imperial College (DIC), London, UK in 1993. He received his PhD degree in Electrical Engineering from City University London, London, UK in 2007. He also obtained a Postgraduate Certificate in Learning and Teaching from the University of West London, London, UK in 2009.

He worked as a Teaching Assistant, Research Assistant, Research Associated, and Postdoctoral Research Fellow at City University London, London, UK from 1998. He worked as a Senior Lecturer and Programme Leader at the University of West London from 2008. He works as a Faculty and Programme Leader of Biomedical Engineering at the Faculty of Engineering Technology and Science at the Higher Colleges of Technology, Dubai, UAE. He authored 30 publications published in international peer-reviewed journals and conferences. Research interests include detection and classification of ECG signals including abnormalities as well as transabdominal fetal ECG signal during labour, adaptive filtering, non-linear and non-Gaussian signals in non-stationary environment, neural networks, higher-order statistics, and chaos and fractal signals.

Dr. Zgallai is a Fellow of the Higher Education Academy in the UK, FHEA. He is a member of the IET, member of the IEEE, and an Associate Member of the Institute of Knowledge Transfer in the UK. He serves in many academic committees, including accreditations, he has been an external examiner at international HE institutions.

AUTHOR INDEX