Computer Science & Information Technology 46

Jan Zizka
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

Fourth International Conference on Advanced Information Technologies
and Applications (ICAITA 2015)
Dubai, UAE, November 06~07, 2015

AIRCC Publishing Corporation

Volume Editors

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

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-43-4
DOI : 10.5121/csit.2015.51501 - 10.5121/csit.2015.51520

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

Preface

The Fourth International Conference on Advanced Information Technologies and Applications (ICAITA 2015) was held in Dubai, UAE, during November 06~07, 2015. The Fourth International Conference on Soft Computing, Artificial Intelligence and Applications (SAI 2015), The Fourth International Conference on Data Mining & Knowledge Management Process (CDKP 2015), The Second International Conference on Signal and Image Processing (Signal 2015) and The International Conference on Networks and Communications (NCO 2015) were collocated with the ICAITA-2015. 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 ICAITA-2015, SAI-2015, CDKP-2015, Signal-2015, NCO-2015 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, ICAITA-2015, SAI-2015, CDKP-2015, Signal-2015, NCO-2015 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 ICAITA-2015, SAI-2015, CDKP-2015, Signal-2015, NCO-2015.

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.

Jan Zizka
Dhinaharan Nagamalai

Organization

General Chair

Natarajan Meghanathan	Jackson State University, USA
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia

Program Committee Members

A.Arokiasamy	Eastern Mediterranean University , Cyprus
Abdallah Rhattoy	Moulay Ismail University, Morocco
Abdelkrim Khireddine	University of Bejaia, Algeria
Abdolreza Hatamlou	Islamic Azad University, Iran
Abdul kadir Ozcan	The American university, Cyprus
Abe Zeid	Northeastern University, USA
Abraham Sanchez Lopez	Autonomous University of Puebla, Mexico
Adamu Murtala Zungeru	Federal University Oye, Nigeria
Adnan Albar	King AbdulAziz University, Saudi Arabia
Adnan H. Ali	Institute of Technolgy, Iraq
Afshari	Islamic Azad University, Iran
Ahmad Lotfi	Nottingham Trent University, United Kingdom
Ahmed samir	Ain-Shams University, Egypt
Ahmed Y. Nada	Al-Quds University, Palestine
Alaa Hussein Al-hamami	Amman Arab University, Jordan
Ali Abid D. Al-Zuky	Mustansiriyah University, Iraq
Ali Chaabani	National School of Engineering, Tunisia
Ali Dorri	Islamic Azad University, Iran
Ali El-Zaart	Beirut Arab University, Lebanon
Almir Pereira Guimaraes	Federal University of Alagoas, Brazil
Ali Hussein	Alexandria University, Egypt
Ali Poorebrahimi	Islamic Azad University, Iran.
Ali Riza YILDIZ	Bursa Technical University, Turkey
Ali Zaart	Beirut Arab University, Lebanon
Ankit Chaudhary	Truman State University, USA
Apai	Universiti Malaysia Perlis, Malaysia
Arifa Ferdousi	Varendra University, Bangladesh
Ashraf A. Shahin	Cairo University, Egypt
Ayad Ghany Ismaeel	Technical Engineering College,Iraq
Baghdad Atmani	University of Oran, Algeria
Barbaros Preveze	Çankaya University, Turkey
Bong-Han Kim	Chongju University, South Korea
ChanChristine	University of regina, Canada
Chih-Lin Hu	National Central University, Taiwan
Chin-Chih Chang	Chung Hua University , Taiwan
Dammak Nouha	Miracl laboratory, Tunisia

Daqiang Zhang	Nanjing Normal University, China
David W Deeds	Shingu College, South Korea
Derya Birant	Dokuz Eylul University, Turkey
Emilio Jimenez Macias	University of La Rioja, Spain
Faiyaz Ahmad	Integral University, Lucknow
Farhan	University of Indonesia, Indonesia
Fatih Korkmaz	Çankiri Karatekin University, Turkey
Fernando Bobillo	University of Zaragoza, Spain
Geuk Lee	Hannam University,South Korea
Gh.A.Montazer	Tarbiat Modares University, Iran
Govindavaram Madhusri	Kakatiya University, India
Grienggrai Rajchakit	Maejo University, Thailand
Gullanar M Hadi	Salahaddin University, Hawler, Iraq
Habib Rasi	Shiraz University of Technology, Iran
Hacene Belhadef	University of Constantine 2, Algeria
Hamdi hassen	MIRACL Laboratory, Tunisia
Hamdi M	National Engineering School of Tunis, Tunisia
Hamid Mcheick	Universite du Quebec a Chicoutimi,Canada
Hamid Taghavifar	Urmia University, Iran
Hanan Salam	University of Pierre and Marie Curie, France
Hangwei	Western Reserve University, USA
Hassini Noureddine	University of Oran , Algeria
Hesham Farouk	Electronics Research Institute, Egypt
Hossein Jadidoleslami	MUT University, Iran
I-Ching Hsu	National Formosa University, Taiwan
Ing. habil. Natasa Zivic	University of Siegen, Germany
Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University, Egypt
Israashaker Alani	Gaziantep University, Turkey
Iwan Adhicandra	University of Pisa, Italy
Jacques Epounde Ngalle	Robert Morris University, USA
Jae-Kwang Lee	Hannam University, South Korea
Jeremy (Zheng) Li	University of Bridgeport, USA
Jose Raniery	University of Sao Paulo, Brazil
Keneilwe Zuva	University of Botswana, Botswana
Koczy T.Laszlo	Budapest University of Technology, Hungary
Laudson Souza	Professor of Integrated Faculties of Patos, Brazil
M.Hamadouche	Universite Saad Dahlab de Blida, Algeria
Mahdi Mazinani	Islamic Azad University, Iran
Malek	Jadara University, Jordan
Marjan Mernik	University of Alabama at Birmingham, USA
Meachikh	University of Tlemcen, Algeria
Mohamed Elboukhari	Mohamed I University, Morocco
Nabila labraoui	University of Abou Bekr Belkaid, Algeria
Narasimha Inukollu	University of Houston, USA
Natarajan Meghanathan	Jackson State University, USA
Nguyen Dinh	University of Science, VIETNAM
Nor Aniza Abdullah	University of Malaya, Malaysia
Noureddine Bouhmala	Buskerud and Vestfold University, Norway

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 Fourth International Conference on Advanced Information Technologies and Applications (ICAITA 2015)

The Fourth International Conference on Soft Computing, Artificial Intelligence and Applications (SAI 2015)

The Fourth International Conference on Data Mining & Knowledge Management Process (CDKP 2015)

The Second International Conference on Signal and Image Processing (Signal 2015)

The International Conference on Networks and Communications (NCO 2015)

USE OF EIGENVECTOR CENTRALITY TO DETECT GRAPH ISOMORPHISM

Natarajan Meghanathan

Jackson State University, 1400 Lynch St, Jackson, MS, USA
natarajan.meghanathan@jsums.edu

ABSTRACT

Graph Isomorphism is one of the classical problems of graph theory for which no deterministic polynomial-time algorithm is currently known, but has been neither proven to be NP-complete. Several heuristic algorithms have been proposed to determine whether or not two graphs are isomorphic (i.e., structurally the same). In this research, we propose to use the sequence (either the non-decreasing or non-increasing order) of eigenvector centrality (EVC) values of the vertices of two graphs as a precursor step to decide whether or not to further conduct tests for graph isomorphism. The eigenvector centrality of a vertex in a graph is a measure of the degree of the vertex as well as the degrees of its neighbors. We hypothesize that if the non-increasing (or non-decreasing) order of listings of the EVC values of the vertices of two test graphs are not the same, then the two graphs are not isomorphic. If two test graphs have an identical non-increasing order of the EVC sequence, then they are declared to be potentially isomorphic and confirmed through additional heuristics. We test our hypothesis on random graphs (generated according to the Erdos-Renyi model) and we observe the hypothesis to be indeed true: graph pairs that have the same sequence of non-increasing order of EVC values have been confirmed to be isomorphic using the well-known Nauty software.

KEYWORDS

Graph Isomorphism, Degree, Eigenvector Centrality, Random Graphs, Precursor Step

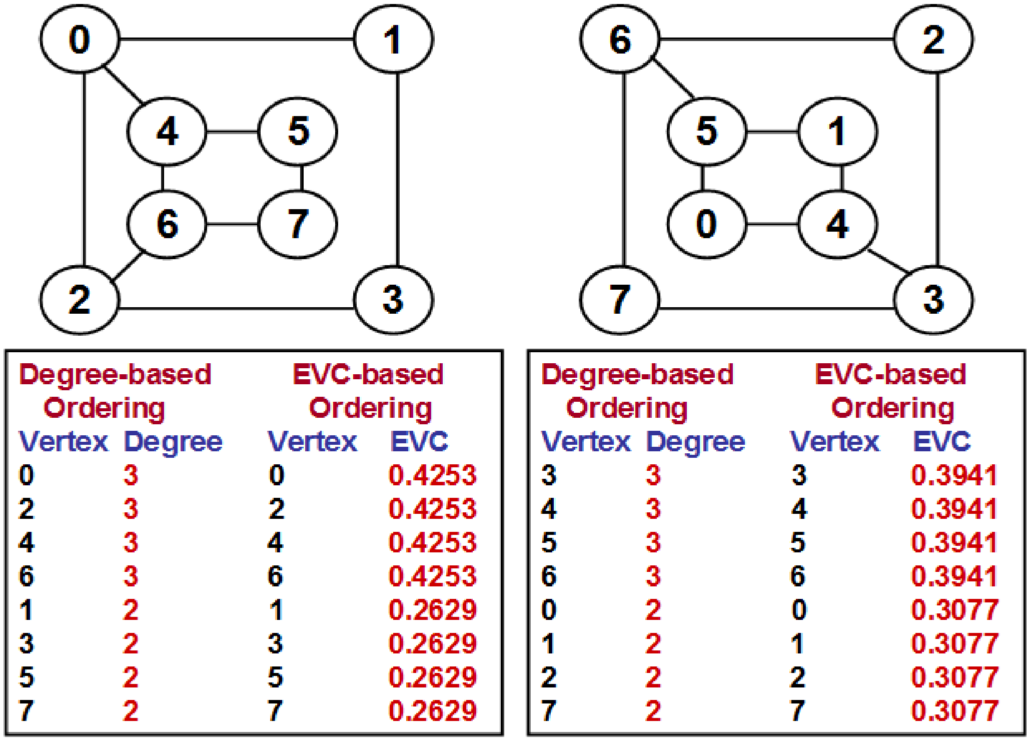
1. INTRODUCTION

Graph isomorphism is one of the classical problems of graph theory for which there exist no deterministic polynomial-time algorithm and at the same time the problem has not been yet proven to be NP-complete. Given two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ - where V_1 and E_1 are the sets of vertices and edges of G_1 and V_2 and E_2 are the sets of vertices and edges of G_2 - we say the two graphs are isomorphic, if the two graphs are structurally the same. In other words, two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ are isomorphic [1] if and only if we can find a bijective mapping f of the vertices of G_1 and G_2, such that $\forall\, v \in V_1, f(v) \in V_2$ and $\forall\, (u, v) \in E_1, (f(u), f(v)) \in E_2$. As the problem belongs to the class NP, several heuristics (e.g., [7-9]) have been proposed to determine whether any two graphs G_1 and G_2 are isomorphic or not. The bane of these heuristics is that they are too time-consuming for large graphs and could lead to identifying several false positives (i.e., concluding a pair of two non-isomorphic graphs as isomorphic).

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 01–09, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51501

To minimize the computation time, the test graphs (graphs that are to be tested for isomorphism) are subject to one or more precursor steps (pre-processing routines) that could categorically discard certain pair of graphs as non-isomorphic (without the need for validating further using any time-consuming heuristic). For two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ to be isomorphic, a basic requirement is that the two graphs should have the same number of vertices and similarly the same number of edges. That is, if $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ are to be isomorphic, then it implies $|V_1| = |V_2|$ and $|E_1| = |E_2|$. If $|V_1| \neq |V_2|$ and/or $|E_1| \neq |E_2|$, then we can categorically say that G_1 and G_2 are not isomorphic and the two graphs need not be processed further through any time-consuming heuristics to test for isomorphism.

In addition to checking for the number of vertices and edges, one of the common precursor steps to test for graph isomorphism is to determine the degree of the vertices of the two graphs that are to be tested for isomorphism and check if a non-increasing order (or a non-decreasing order; we will follow a convention of sorting in a non-increasing order) of the degrees of the vertices of the two graphs is the same. If the non-increasing order of the degree sequence of two graphs G_1 and G_2 are not the same, then the two graphs can be categorically ruled out from being isomorphic. If two graphs are isomorphic, then identical degree sequence of the vertices in a particular sorted order is a necessity. However as shown in Figure 1, it is possible that two graphs could have the same degree sequence in a particular sorted order, but need not be isomorphic [2]. Though very time-efficient, the degree sequence-based precursor step to test for graph isomorphism is typically considered to be erratic and not reliable (leading to false positives), especially while testing for isomorphism among graphs with a smaller number of vertices (like the example in Figure 1).



Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
0	3	0	0.4253
2	3	2	0.4253
4	3	4	0.4253
6	3	6	0.4253
1	2	1	0.2629
3	2	3	0.2629
5	2	5	0.2629
7	2	7	0.2629

Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
3	3	3	0.3941
4	3	4	0.3941
5	3	5	0.3941
6	3	6	0.3941
0	2	0	0.3077
1	2	1	0.3077
2	2	2	0.3077
7	2	7	0.3077

Figure 1: Example for Two Non-Isomorphic Graphs with the Same Degree Sequence, but Different Eigenvector Centrality (EVC) Sequence

Centrality metrics are one of the commonly used quantitative measures to rank the vertices of a graph based on the topological structure of the graph [3]. Degree centrality is one of the primitive and typically used centrality metrics for complex network analysis; but, in addition to the weakness illustrated in Figure 1 and explained in the previous paragraph, it is also evident from Figure 1 that degree centrality-based ranking of the vertices could result in ties (i.e., the technique has weak discrimination power) among vertices having the same degree (as the degree centrality values are integers) and it may not be possible to unambiguously rank the vertices; for graphs of any size, it is likely that more than one vertex may have the same degree (ties). Eigenvector centrality (EVC) is a well-known centrality measure in the area of complex networks [4]. The EVC of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors (calculations of EVC values is discussed in Section 2). For example: if two vertices *X* and *Y* have degree 3, but if all the three neighbors of *X* have a degree 2 and if at least one of the neighbors of *Y* have degree greater than 2 and others have degree at least 2, then the EVC of *Y* is guaranteed to be greater than the EVC of *X*. In general, the EVC of a vertex not only depends on the degree of the vertex, but also on the degree of its neighbors. For a connected graph, the EVC values of the vertices are positive real numbers in the range (0...1) and are more likely to be different from each other, contributing to the scenario of unambiguous ranking of the vertices as much as possible (the EVC technique has a relatively stronger discrimination power compared to the degree-based technique).

With respect to Figure 1, we notice that the non-increasing order listings of the EVC values of the vertices for the two graphs are not the same. The discrepancy is obvious in the largest EVC value of the two sequences itself. The largest EVC value for a vertex in the first graph is 0.4253 and the largest EVC value for a vertex in the second graph is 0.3941. The example in Figure 1 is a motivation for our hypothesis to use the EVC values as the basis for deciding whether or not two graphs could be isomorphic.

The rest of the paper is organized as follows: Section 2 explains the procedure to determine the Eigenvector Centrality (EVC) values of the vertices. In Section 3, we propose the use of the Eigenvector Centrality (EVC) measure as the basis of the precursor step to determine whether or not two graphs are isomorphic. In Section 4, we test our hypothesis on random network graphs (generated according to the Erdos-Renyi model [5]) with regards to the application of the EVC measure for detecting isomorphism among graphs. Section 5 discusses related work. Section 6 concludes the paper. Throughout the paper, the terms 'node' and 'vertex' as well as 'edge' and 'link' are used interchangeably. They mean the same.

2. EIGENVECTOR CENTRALITY

The Eigenvector Centrality (EVC) of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors. The EVC of the vertices in a network graph is the principal eigenvector of the adjacency matrix of the graph. The principal eigenvector has an entry for each of the *n*-vertices of the graph. The larger the value of this entry for a vertex, the higher is its ranking with respect to EVC. We illustrate the use of the Power-iteration method [6] (see example in Figure 2) to efficiently calculate the principal eigenvector for the adjacency matrix of a graph. The eigenvector X_{i+1} of a network graph at the end of the $(i+1)^{th}$ iteration is given by: $X_{i+1} = \frac{AX_i}{\|AX_i\|}$, where $\|AX_i\|$ is the normalized value of the product of the adjacency matrix A of a given graph and the tentative eigenvector X_i at the end of iteration *i*. The initial value of X_i is the transpose of [1, 1, ..., 1], a column vector of all 1s, where the number of 1s correspond to the number of vertices in the graph. We continue the iterations until the normalized value $||AX_{i+1}||$ converges to that of the normalized value $||AX_i||$. The value of the column vector X_i at this juncture is declared the Eigenvector centrality of the graph; the entries corresponding to the individual rows in X_i represent the Eigenvector centrality of the vertices of the graph. The converged normalized value of the Eigenvector is referred to as the Spectral radius.

As can be seen in the example of Figure 2, the EVC of a vertex is a function of both its degree as well as the degree of its neighbors. For instance, we see that both vertices 2 and 4 have the same degree (3); however, vertex 4 is connected to three vertices that have a high degree (3); whereas vertex 2 is connected to two vertices that have a relatively low degree (of degree 2); hence, the EVC of vertex 4 is larger than that of vertex 2. As can be seen in the example of Figure 2, the EVC values of the vertices are more likely to be distinct and could be a better measure for unambiguously ranking the vertices of a network graph.

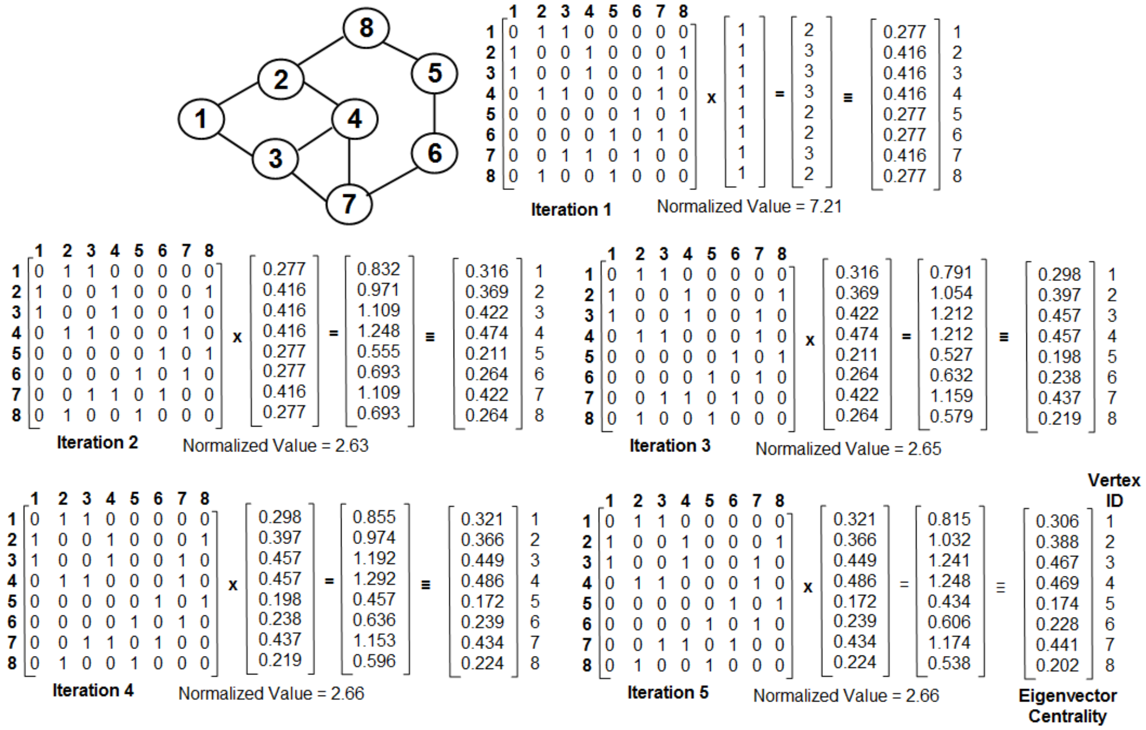
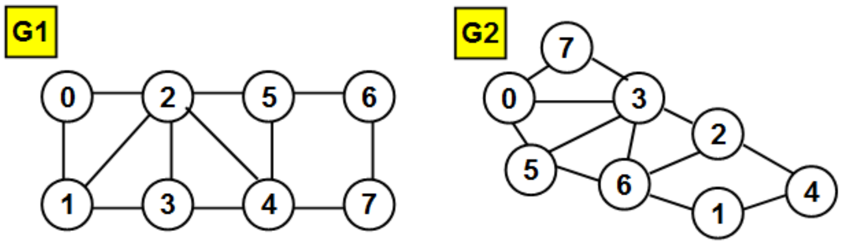


Figure 2: Example to Illustrate the Computation of Eigenvector Centrality (EVC) of the Vertices using the Power-Iteration Method

The number of iterations needed for the normalized value of the eigenvector to converge is anticipated to be less than or equal to the number of vertices in the graph [6]. Each iteration of the power-iteration method requires $\Theta(V^2)$ multiplications, where V is the number of vertices in the graph. With a maximum of V iterations expected, the overall time complexity of the algorithm to determine the Eigenvector Centrality of the vertices of a graph of V vertices is $\Theta(V^3)$.

3. HYPOTHESIS

Our hypothesis is that if a non-increasing order of listings of the EVC values of the vertices for two graphs G_1 and G_2 are not identical, then the two graphs are not isomorphic. If the non-increasing sequence of EVC values for the two graphs is identical, we declare the two graphs to be potentially isomorphic and subject them to further tests for isomorphism (for confirmation). Thus, the technique of listing the EVC sequence of the vertices (in a non-increasing order) could be used as an effective precursor step before subjecting the graphs to any time-consuming heuristic for graph isomorphism. As the EVC values of the vertices in any random graph are more likely to be unique, this test would also help us to extract a mapping of the vertices between two graphs that have been identified to be potentially isomorphic and make it more easy for the time-consuming complex heuristics to test for isomorphism. We illustrate our hypothesis using an example in Figure 3. From the example, it is very obvious that if two graphs have an identical non-increasing order listing of the EVC sequence, they should have identical non-increasing order listing of the degree sequence; but, not vice-versa (refer example in Figure 1). If two graphs have a different non-increasing order of degree sequence, they cannot have the same non-increasing order of EVC sequence and we do not need to compute the EVC values.



Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
2	5	2	0.5364
4	4	4	0.4321
1	3	3	0.3974
3	3	1	0.3596
5	3	5	0.3355
0	2	0	0.2681
6	2	7	0.1749
7	2	6	0.1527

Degree-based Ordering		EVC-based Ordering	
Vertex	Degree	Vertex	EVC
3	5	3	0.5364
6	4	6	0.4321
0	3	5	0.3974
2	3	0	0.3596
5	3	2	0.3355
1	2	7	0.2681
4	2	1	0.1749
7	2	4	0.1527

Mapping of Vertices	
G1	G2
2	3
4	6
3	5
1	0
5	2
0	7
7	1
6	4

Figure 3: Illustration of the Hypothesis: Eigenvector Centrality (EVC) to Decide Graph Isomorphism

We notice from Figure 3 that the vertices corresponding to the non-increasing order of the EVC values in both the graphs could be uniquely mapped to each other on a one-to-one basis (bijective mapping). On the other hand, the non-increasing order of the degree sequence of the vertices merely facilitates us to group the vertices into different equivalence classes (all vertices of the same degree in both the graphs are said to be equivalent to each other); but, one could not arrive at a unique one-to-one mapping of the vertices that corresponds to the structure of the two graphs. We thus hypothesize that the EVC approach could not only help us to determine whether or not two graphs are isomorphic, it also facilitates us to potentially arrive at a unique one-to-one mapping of the vertices in the corresponding two graphs and feed such a mapping as input to any heuristic that is used to confirm whether two graphs that have been identified to be possibly isomorphic (using the EVC approach) are indeed isomorphic.

4. SIMULATIONS

We tested our hypothesis by conducting extensive simulations on random network graphs generated according to the Erdos-Renyi model [5]. According to this model, the network has N nodes and the probability of a link between any two nodes is p_{link}. For any pair of vertices u and v, we generate a random number in the range [0...1] and if the random number is less than p_{link}, there is a link between the two vertices u and v; otherwise, not. We constructed random networks of N = 10 nodes with p_{link} values of 0.2 to 0.8 (in increments of 0.1). We constructed a suite of 1000 networks for each value of p_{link}. We chose a smaller value for the number of nodes as we did not observe any pair of isomorphic graphs in a suite of 1000 graphs created with N = 100 nodes for any p_{link} value. Even for networks of N = 10 nodes, there is a high chance of observing pairs of isomorphic graphs only under low or high values of p_{link}. For p_{link} values of 0.2 and 0.3, the pairs of isomorphic graphs observed were typically trees (graphs without any cycles) that have the minimal number of edges to keep all the nodes connected. As we increase the number of links in the networks, the chances of finding any two distinct isomorphic random graphs get extremely small. On the other hand, for p_{link} values of 0.7 and 0.8, the isomorphic graphs were observed to be close to complete graphs (with only one or two missing links per node from becoming a complete graph).

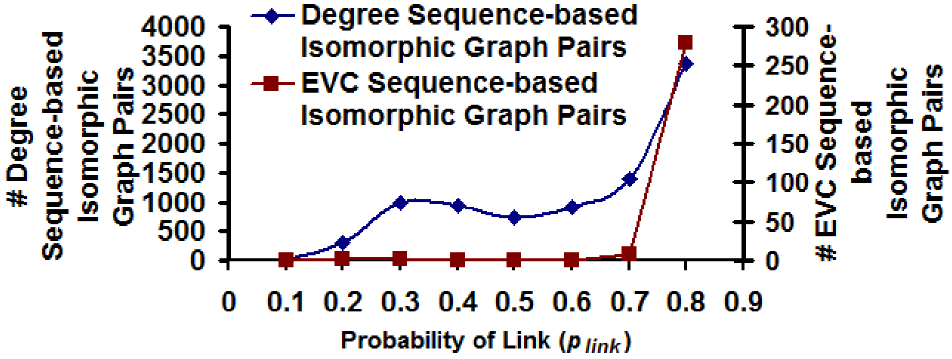


Figure 4: Number of Isomorphic Random Graph Pairs: Degree Sequence vs. EVC Sequence Approach

The success of the hypothesis is evaluated by determining the number of pairs of isomorphic graphs identified based on the non-increasing order of the EVC sequence vis-a-vis the degree sequence. As mentioned earlier, if two graphs are isomorphic, then the non-increasing order of listing of the EVC values of the vertices has to be identical (as the two graphs are essentially the same, with just the vertices labeled differently). This implies that if the non-increasing order of listing of the EVC values of the vertices for a pair of graphs G_1 and G_2 are not identical, we need not further subject the two graphs to any other heuristic test for isomorphism. If two graphs are identified to be potentially isomorphic based on the EVC sequence, we further processed those two graphs using the Nauty software [7] and confirmed that the two graphs are indeed isomorphic to each other. We did not observe any false positives with the EVC approach. The Nauty software [7] is the world's fastest testing software (available at: http://www3.cs.stonybrook.edu/ ~algorith/implement/nauty/implement.shtml) to detect graph isomorphism.

Figure 4 illustrates the number of graph pairs that have been identified to be potentially isomorphic on the basis of the EVC sequence approach vis-a-vis the degree sequence approach. We observe that even with the degree sequence approach, for moderate p_{link} values (0.4-0.5), the number of graph pairs identified to be potentially isomorphic decreases from that observed for low-moderate p_{link} value of 0.3. As we further increase the p_{link} value, the number of graph pairs identified to be potentially isomorphic increases significantly with both the degree sequence and EVC sequence-based approach, and the EVC sequence-based approach identifies a significantly larger number of these graph pairs (that are already identified to be potentially isomorphic based on the degree sequence) to be indeed potentially isomorphic and this is further reconfirmed through the Nauty software. For low-moderate p_{link} values, we observe the degree sequence-based approach to identify an increasingly larger number of graph pairs to be potentially isomorphic, but they were observed to be indeed not isomorphic on the basis of the EVC sequence approach as well as when tested using the Nauty software. This vindicates our earlier assertion (in Section 1) that the degree sequence-based precursor step is prone to incurring a larger number of false positives (i.e., erratically identifying graph pairs as isomorphic when they are indeed not isomorphic).

5. RELATED WORK

Though centrality measures have been widely used for problems related to complex network analysis [3], the degree centrality measure is the only common and most directly used centrality measure to test for graph isomorphism [1]. The other commonly used centrality-based precursor step to test for the isomorphism of two or more graphs is to find the shortest path vector for each vertex in the test graphs and evaluate the similarity of the shortest path matrix (an ensemble of the shortest path vectors of the constituent vertices) of the test graphs. Since the one-to-one mapping between the vertices of the test graphs is not known a priori, one would need a time-efficient algorithm to compare the columns (shortest path vectors) of two matrices for similarity between the columns. The closeness centrality measure [3] is the centrality measure that matches to the above precursor step. Both the degree and closeness centrality measures have an inherent weakness of incurring only integer values (contributing to their poor discrimination of the vertices) and it is quite possible that two or more vertices have the same integer value under either of these centrality measures and one would not be able to obtain a distinct ranking of the vertices (i.e., unique values of the centrality scores) to detect for graph isomorphism. The eigenvector centrality measure incurs real numbers as values in the range (0...1) and has a much higher chance of incurring distinct values for each of the vertices of a graph. Though there could be scenarios where two or more vertices have the same EVC value, a non-increasing or non-decreasing order listing the EVC values of the vertices of two different graphs is more likely to be different from each other if the two graphs are non-isomorphic. As the complexity of the graph topology increases (as the number of vertices and edges increases), we observed it to be extremely difficult to generate two random graphs that have the same sequence (say in the non-increasing order) of EVC values for the vertices and be isomorphic.

As mentioned earlier, graph isomorphism is one of the classical problems of graph theory that has not been yet proven to be NP-complete, but there does not exist a deterministic polynomial time algorithm either. Many heuristics have been proposed to solve the graph isomorphism problem (e.g., Nauty [7], Ullmann algorithm [8] and VF2 [9]), but all of them take an exponential time at the worst case as most of them take the approach of progressively searching for all possible matching between the vertices of the test graphs. To reduce the search complexity, the heuristics could use precursor steps like checking for identical degree sequence for the vertices of the test graphs. It would be preferable to use precursor steps that contribute to fewer false positives, if not none. This is where our proposed approach of using the eigenvector centrality (EVC) fits the bill. We observe from the simulations that all the graphs identified to be isomorphic (using the EVC approach) are indeed isomorphic. Thus, the EVC sequence-based listing of the vertices could be rather used as an effective precursor step to rule out graph pairs that are guaranteed to be not isomorphic, especially when used with the more recently developed time-efficient heuristics that effectively prune the search space (e.g., the parameterized matching [10] algorithm).

The eigenvector centrality (EVC) measure falls under a broad category of measures called "graph invariants" that have been extensively investigated in discrete mathematics [11-12], structural chemistry [13-14] and computer science [15]. These graph invariants can be classified to be either global (e.g., Randic index [16]) or local (e.g., vertex complexity [17]) as well as be either information-theoretic (statistical quantities) [18-19] or non-information-theoretic indices [20]. With the objective of reducing the run-time complexity of the heuristics for graph isomorphism, weaker but time-efficient precursor tests (measures with poor discrimination power like the degree sequence) were rather commonly used. Sometimes, a suite of such simplistic graph invariants were used [21] and test graphs observed to be potentially isomorphic based on each of these invariants were considered for further analysis with a complex heuristic. The discrimination power of the weaker graph invariants also vary with the type of graphs studied [21]. To the best of our knowledge, the discrimination power of the more complex graph invariants - especially those based on the spectral characteristics of a graph (like that of the Eigenvector Centrality), is yet to be analyzed. Ours is the first effort in this direction.

6. CONCLUSIONS

The high-level contribution of this paper is the proposal to use the Eigenvector Centrality (EVC) measure to detect isomorphism among two or more graphs. We propose that if the non-increasing order (or non-decreasing order) of listing the EVC values of the vertices of the test graphs are not identical, then the test graphs are not isomorphic and need not be further processed by any time-consuming heuristic to detect graph isomorphism. This implies that if two or more graphs are isomorphic to each other, their EVC values written in the non-increasing order must be identical. We test our hypothesis on a suite of random network graphs generated with different values for the probability of link and observed the EVC approach to be effective: there are no false positives, unlike the degree sequence based approach. The graph pairs that are observed to have an identical EVC sequence are confirmed to be indeed isomorphic using the Nauty graph isomorphism detection software. We also observe it to be extremely difficult to generate isomorphic random graphs under moderate values for the probability of link (0.4-0.6); it is rather relatively more easy to generate isomorphic random graphs that are either trees (created when the probability of link values are low: 0.2-0.3) or close to complete graphs (created when the probability of link values are high: 0.7-0.8).

REFERENCES

[1] R. Diestel, Graph Theory (Graduate Texts in Mathematics), Springer, 4th edition, October 2010.

[2] S. Pemmaraju and S. Skiena, Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Cambridge University Press, December 2003.

[3] M. Newman, Networks: An Introduction, 1st ed., Oxford University Press, May 2010.

[4] S. P. Borgatti and M. G. Everett, "A Graph-Theoretic Perspective on Centrality," Social Networks, vol. 28, no. 4, pp. 466-484, October 2006.

[5] P. Erdos and A. Renyi, "On Random Graphs. I," Publicationes Mathemticae, vol. 6, pp. 290-297, 1959.

[6] G. Strang, Linear Algebra and its Applications, Brooks Cole, 4th edition, July 2005.

[7] B. D. McKay, Nauty User's Guide (version 1.5), Technical Report, TR-CS-90-02, Department of Computer Science, Australian National University, 1990.

[8] J. R. Ullman, "An Algorithm for Subgraph Isomorphism," Journal of the ACM, vol. 23, no. 1, pp. 31-42, January 1976.

[9] L. P. Cordella, P. Foggia, C. Sansone and M. Vento, "A (Sub)graph Isomorphism Algorithm for Matching Large Graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 10, pp. 1367-1372, October 2004.

[10] J. Mendivelso, S. Kim, S. Elnikety, Y. He, S-W. Hwang and Y. Pinzon, "Solving Graph Isomorphism using Parameterized Matching," Proceedings of the 20th International Symposium on String Processing and Information Retrieval, pp. 230-242, Jerusalem, Israel, October 2013.

[11] F. Harary, Graph Theory, Westview Press, 1st edition, October 1994.

[12] A. Lewis, "The Convex Analysis of Unitarily Invariant Matrix Functions," Journal of Convex Analysis, vol. 2, no. 1-2, pp. 173-183, 1995.

[13] M. V. Diudea, I. Gutman and L. Jantschi, Molecular Topology, Nova Publishing, New York, NY, USA, June 2001.

[14] X. Liu and D. J. Klein, "The Graph Isomorphism Problem," Journal of Computational Chemistry, vol. 12, no. 10, pp. 1243-1251, December 1991.

[15] B. D. McKay, "Graph Isomorphisms," Congress Numerantium, vol. 30, pp. 45-87, 1981.

[16] M. Randic, "On Characterization of Molecular Branching," Journal of American Chemical Society, vol. 97, no. 23, pp. 6609-6615, November 1975.

[17] C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy and S. C. Basak, "Discrimination of Isomeric Structures using Information Theoretic Topological Indices," Journal of Computational Chemistry, vol. 5, no. 6, pp. 581-588, December 1984.

[18] M. Dehmer and A. Mowshowitz, "A History of Graph Entropy Measures," Information Sciences,vol. 181, no. 1, pp. 57-78, January 2011.

[19] R. V. Sole and S. Valverde, "Information Theory of Complex Networks: On Evolution and Architectural Constraints," Lecture Notes in Physics, vol. 650, pp. 189-207, August 2004.

[20] A. Mehler, A. Lucking and P. Weib, "A Network Model of Interpersonal Alignment in Dialog," Entropy, vol. 12, no. 6, pp. 1440-1483, 2010.

[21] M. Dehmer, M. Grabner, A. Mowshowitz and F. Emmert-Streib, "An Efficient Heuristic Approach to Detecting Graph Isomorphism based on Combinations of Highly Discriminating Invariants," Advances in Computational Mathematics, vol. 39, no. 12, pp. 311-325, August 2013.

AUTHOR

Dr. Natarajan Meghanathan is a tenured Full Professor of Computer Science at Jackson State University, USA. His areas of research interests are Network Science and Graph Theory, Wireless Ad hoc Networks and Sensor Networks, Cyber Security and Machine Learning. He has published more than 150 peer-reviewed articles and obtained grants from several federal agencies. He serves as the editor-in-chief of three international journals as well as serves in the organizing committees of several international conferences.

INTENTIONAL BLANK

ANALYSIS OF LEXICO-SYNTACTIC PATTERNS FOR ANTONYM PAIR EXTRACTION FROM A TURKISH CORPUS

Gürkan Şahin[1], Banu Diri[1] and Tuğba Yıldız[2]

[1]Faculty of Electrical-Electronic, Department of Computer Engineering
Yıdız Technical University, İstanbul, Turkey
{gurkans,banu}@ce.yildiz.edu.tr
[2]Faculty of Engineering and Natural Sciences,
Department of Computer Engineering
İstanbul Bilgi University, İstanbul, Turkey
tdalyan@bilgi.edu.tr

ABSTRACT

Extraction of semantic relations from various sources such as corpus, web pages, dictionary definitions etc. is one of the most important issue in study of Natural Language Processing (NLP). Various methods have been used to extract semantic relation from various sources. Pattern-based approach is one of the most popular method among them. In this study, we propose a model to extract antonym pairs from Turkish corpus automatically. Using a set of seeds, we automatically extract lexico-syntactic patterns (LSPs) for antonym relation from corpus. Reliability score is calculated for each pattern. The most reliable patterns are used to generate new antonym pairs. Study conduct on only adjective-adjective and noun-noun pairs. Noun and adjective target words are used to measure success of method and candidate antonyms are generated using reliable patterns. For each antonym pair consisting of candidate antonym and target word, antonym score is calculated. Pairs that have a certain score are assigned to antonym pair. The proposed method shows good performance with 77.2% average accuracy.

KEYWORDS

Natural Language Processing, Semantic relations, Antonym, Pattern-based approach

1. INTRODUCTION

Extraction of semantic relation pairs from corpus is one of the most popular topic in NLP. Hyponymy, hypernymy, meronymy, holonymy, synonymy, antonymy etc. can be given to example of semantic relations.

Several resources are used to acquire semantic relations. WordNet [1] is the one of the important sources for semantic relations. WordNet is a lexical database for English and consists of so many words and links among these words. Since each word is represented as synonym words called

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 11–17, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51502

synsets in WordNet, it can be said that main relations of WordNet is synonymy. Apart from synonymy relation, words are connected each other via semantic relation links like hyponymy, hypernymy, meronymy, holonymy, antonymy etc. Words are collected under four different titles as noun, adjective, verb, adverb, respectively in WordNet.

One of the most important semantic relation in WordNet is antonymy. Antonymy represents contrast sense between two words. In fact, there is no exactly consensus on the definition of the antonymy. According to domain experts, some pairs like good-bad, hot-cold etc. represent good antonymy relation, but some pairs like north-south, woman-man do not exactly represent antonymy. This makes difficult to detect opposite pairs. In addition, studies have shown that synonym and antonym words occur with similar context words. This case also reveals difficulty of distinguishing antonyms from synonyms.

In this study, we propose a pattern-based model to extract antonym pairs from Turkish corpus. Only lexico syntactic patterns are used to find antonym pairs. Noun-noun and adjective-adjective antonym initial seeds are prepared and antonym patterns are extracted using seeds. Patterns having a reliable pattern score are selected to generate new antonym pairs from corpus.

The rest of this paper organized as follows: Section 2 presents related works. Extraction of antonym patterns and extraction of new antonym pairs are explained in Section 3 and Section 4, respectively. Finally, we present experimental results in Section 5.

2. RELATED WORKS

Patterns have been widely used to extract semantic relations from corpus. The most popular pattern-based study was made by Hearst [2] in 1992. Hearst used some patterns like “such X as Y” to extract hyponym words from corpus. In this pattern, X and Y represent hypernym and hyponym words, respectively. After experiments, it has been shown that using some patterns, hyponym words can be extracted from corpus with high accuracy.

Various studies have been conducted on extraction of antonym pairs. Lobanova (2010) [3] prepared some adjective-adjective antonym initial pairs and generated antonym patterns occurring with initial pairs from large Dutch corpus. Lobanova used generated antonym patterns to extract new antonym pairs from corpus. This process repeated iteratively. At each iteration, new antonym patterns were generated by using initial pairs and new antonym pairs were used to extract new antonym patterns again. At each iteration, only reliable antonym pairs and patterns were selected. Thus, sharp accuracy decreasing for generated antonym pairs was prevented.

Turney (2008) [4] used a corpus based supervised classification method to separate antonyms from synonyms. Only patterns obtained from corpus were used as features. Co-occurrence frequency between pair and pattern was used as a feature. Support Vector Machines (SVM) was used as classification algorithm. To measure success of method, English as a second language (ESL) questions were used and 75.0% classification accuracy was obtained for antonym pair classification.

Lin (2003) [5] manually prepared some patterns like “from X to Y”, “either X or Y” to discriminate synonyms from antonyms. It was observed that antonym pairs occur with these two pattern very frequently but synonym pairs occur with these patterns rarely.

Mohammad (2008) [6] developed an unsupervised method using degree of antonym to discriminate antonym pairs. According to definition of degree of antonym, the more a pair has antonym degree, the more the pair represents antonymy. Mohammad used corpus statistical features and antonym dictionary category words together. Over test pairs 80.0% accuracy was obtained for antonym pairs.

For Turkish, there are some studies to extract semantic relation pairs from corpus and dictionary definitions [7], [8]. Hyponym-hypernym [9], [10], meronym-holonym [11] and synonym [12] pairs have been automatically extracted from Turkish corpus. For hyponym-hypernym, meronym-holonym and synonym pairs 83.0%, 75.0% and 80.3% accuracies were obtained, respectively.

Although there are some studies about antonym pair extraction from Turkish dictionary definitions, there is no study using Turkish corpus and antonym corpus patterns. Our main motivation is that there is no such a corpus based study for Turkish before.

3. Extraction of Antonym Patterns From Turkish Corpus

LSPs are widely used to extract antonym relation pairs. In this study, antonym patterns are used to extract antonym pairs. Therefore, we have to generate antonym patterns from corpus. To extract Turkish antonym patterns, following processing steps are applied.

- BOUN web corpus was used [13] as a source. The corpus consists of nearly 10 million sentences and 500 million words (tokens). Firstly, we remove all punctuation and special characters from corpus. Corpus is parsed morphologically by Zemberek Turkish NLP tool [14] and each word in corpus is separated to root, root part-of-speech tag and suffixes. For a given word, Zemberek generates multiple parsing results, but only first parsing result is used. Because our corpus is too big, search process can take a long time. For fast search operations, morphologically parsed corpus is indexed by Apache Lucene 4.2.0 searching tool [15] and index file is used for all corpus search operations.

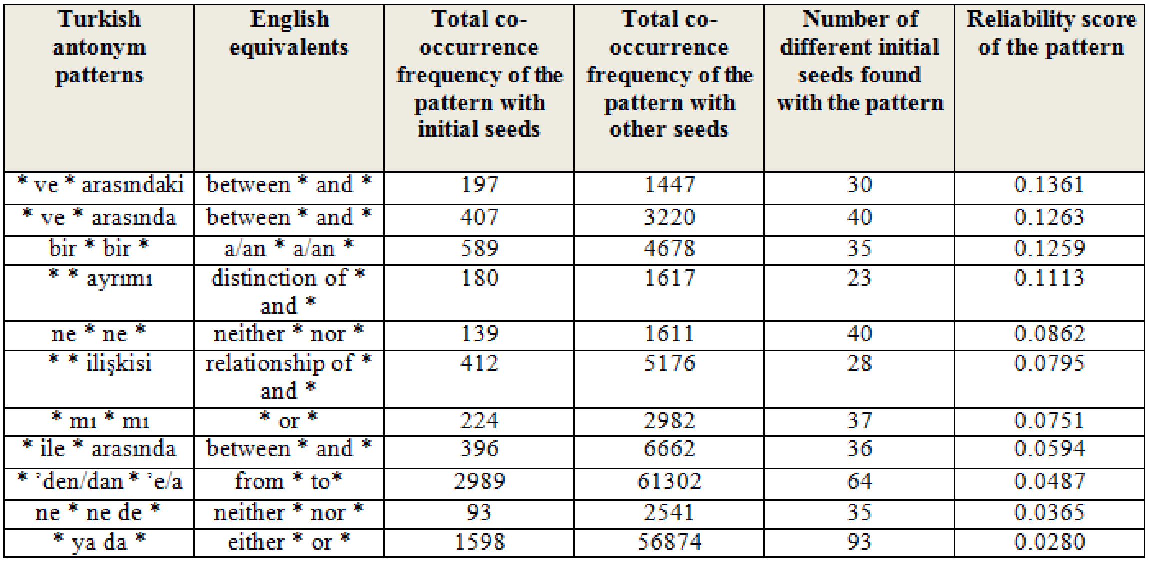
- To find antonym patterns, we generate noun and adjective target words. Antonym equivalents of target antonyms are extracted with using Turkish Antonym Dictionary [16]. 184 antonym pairs called initial seeds are searched in corpus index file and sentences which contain initial seeds are found. In related sentences, initial seeds are replaced with * (wildcard) character. We select patterns having maximum two words between two * characters and others are removed. Thus, we ignore unproductive special patterns.

- Reliability score of each pattern is calculated. To calculate pattern reliability score, we used a formula which is given in equation (1).

$$R_n = \frac{P}{T} \qquad (1)$$

In formula, R_n represents reliability score of pattern n. P is total co-occurrence frequency of pattern n with initial seeds. T represents total co-occurrence frequency of pattern n with other antonym pairs(other seeds) in corpus. Total co-occurrence frequency of pattern with initial seeds is divided by total co-occurrence frequency of pattern with other seeds. Then, reliability score is calculated for each pattern. For example, if pattern X occurs with initial seeds 100 times and occurs with other seeds 10.000 times in corpus, reliability score of X equals 100/10.000 = 0.01. But the reliability score may be misleading. If pattern X occurs with initial seeds 7 times and occurs with other seeds 10 times, reliability score equals 7/10 = 0.7. Although reliability score of X is high, X occurs with initial seeds only 7 times. Because co-occurrence frequency of X with initial seeds is too low, pattern X does not have any importance in terms of productivity and generality. For this reason, we calculate reliability score for patterns occurring with initial seeds more than 50 times and other patterns are ignored. To determine pattern reliability score, number of different initial seeds occurring with a pattern is an important parameter. We can say that the more different initial seeds occur with a pattern, the more the pattern is reliable. We assume that pattern X occurs with initial seeds 100 times, but only occurs with 5 different initial seeds. Likewise, pattern Y occurs with initial seeds 100 times, but occurs with 20 different initial seeds. If total co-occurrence frequency of X and Y with other seeds equals 1000, reliability scores of both patterns equal 100/1000 = 0.1. Although pattern reliability scores of X and Y equal each other, Y pattern occurs with more different initial seeds than X. Hence the pattern Y is more general and productive than X. To tackle this problem, pattern reliability score is calculated for patterns occurring with more than 20 different initial seeds and other patterns are not assessed. After calculating reliability score for each pattern according to two conditions given above, all patterns are sorted according to reliability score. Patterns that have reliability score greater than 0.02 are selected to generate new antonym pairs from corpus. Reliable antonym patterns are given in Table 1.

Table 1. Antonym patterns extracted from corpus using initial seeds



Turkish antonym patterns	English equivalents	Total co-occurrence frequency of the pattern with initial seeds	Total co-occurrence frequency of the pattern with other seeds	Number of different initial seeds found with the pattern	Reliability score of the pattern
* ve * arasındaki	between * and *	197	1447	30	0.1361
* ve * arasında	between * and *	407	3220	40	0.1263
bir * bir *	a/an * a/an *	589	4678	35	0.1259
* * ayrımı	distinction of * and *	180	1617	23	0.1113
ne * ne *	neither * nor *	139	1611	40	0.0862
* * ilişkisi	relationship of * and *	412	5176	28	0.0795
* mı * mı	* or *	224	2982	37	0.0751
* ile * arasında	between * and *	396	6662	36	0.0594
* 'den/dan * 'e/a	from * to*	2989	61302	64	0.0487
ne * ne de *	neither * nor *	93	2541	35	0.0365
* ya da *	either * or *	1598	56874	93	0.0280

4. EXTRACTING NEW ANTONYM PAIRS USING PATTERNS

Using antonym patterns in Table 1, antonym equivalent words are extracted for a given target word. Process steps of new antonym pair extraction are given below.

- Firstly, target words are determined and patterns generated by replacing target word with * characters are searched in corpus. Words corresponding to * characters are extracted as candidates of target word. Although reliable patterns are used, not antonym pairs can occur in these patterns. For this reason, antonym equivalents of given a target word are defined as candidates. In pattern structure, any antonym pairs can show two different sequence like X-Y and Y-X. Thus, given a target word is searched in two different positions and words in different * positions are recorded as candidates. For example, target word "iyi" (good) are searched as;

Turkish patterns	**English equivalents**
# iyi ve * arasındaki	# between good and *
# * ve iyi arasındaki	# between * and good
# ne iyi ne de *	# neither good nor *
# ne * ne de iyi	# neither * nor good

…

- After extracting candidates of target word, antonym score is calculated for each pair consisting of target and a candidate. Pairs having a certain antonym score are assigned to antonym and others are eliminated.

To calculate antonym scores of pairs, we used Lobanova's antonym score formula given in equation (2) [17].

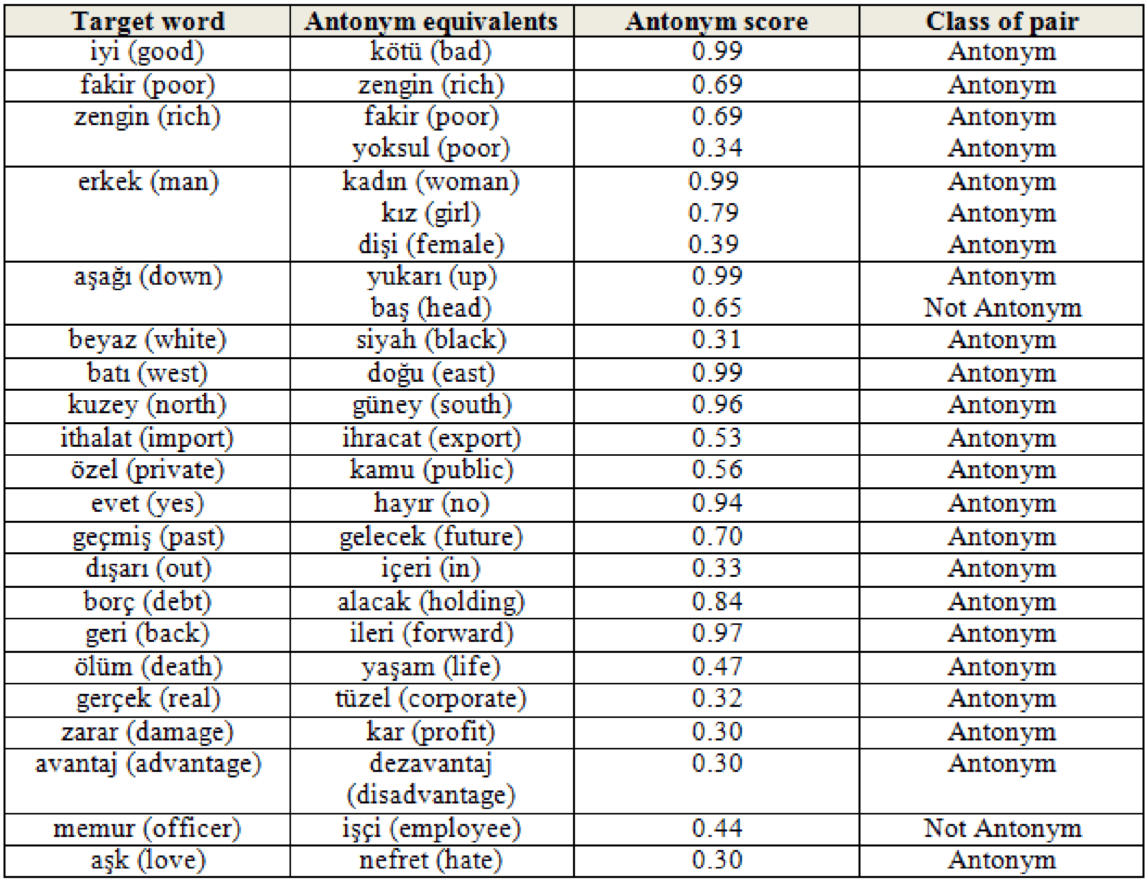
$$P_x = 1 - \prod_{n=1}^{M}\left(1 - \frac{Ck}{Tk}\right)^{Ck} \qquad (2)$$

In formula, P_x represents antonym score for pair x. M is number of reliable pattern and C_k is co-occurrence frequency of pair x with pattern k. T_k represents co-occurrence frequency of pattern k with other seeds in corpus.

5. Experimental Results

To measure success of model, 196 noun and adjective target words are utilized. Target words were searched together with reliable antonym patterns and candidates were extracted. For each pair, antonym score was calculated. After observations, we defined minimum reliable antonym score as 0.3. When the minimum reliable antonym score is defined less than 0.3, it is shown that accuracy of the method falls sharply. For 45 out 196 target words, our method proposed reliable antonym pairs with 77.2% average accuracy. Class of pairs were manually tagged by 3 Turkish native speakers. 21 target words and candidates, english equivalents are given in Table 2.

Table 2. Target words, candidates, antonym scores and pair classes



Target word	Antonym equivalents	Antonym score	Class of pair
iyi (good)	kötü (bad)	0.99	Antonym
fakir (poor)	zengin (rich)	0.69	Antonym
zengin (rich)	fakir (poor) yoksul (poor)	0.69 0.34	Antonym Antonym
erkek (man)	kadın (woman) kız (girl) dişi (female)	0.99 0.79 0.39	Antonym Antonym Antonym
aşağı (down)	yukarı (up) baş (head)	0.99 0.65	Antonym Not Antonym
beyaz (white)	siyah (black)	0.31	Antonym
batı (west)	doğu (east)	0.99	Antonym
kuzey (north)	güney (south)	0.96	Antonym
ithalat (import)	ihracat (export)	0.53	Antonym
özel (private)	kamu (public)	0.56	Antonym
evet (yes)	hayır (no)	0.94	Antonym
geçmiş (past)	gelecek (future)	0.70	Antonym
dışarı (out)	içeri (in)	0.33	Antonym
borç (debt)	alacak (holding)	0.84	Antonym
geri (back)	ileri (forward)	0.97	Antonym
ölüm (death)	yaşam (life)	0.47	Antonym
gerçek (real)	tüzel (corporate)	0.32	Antonym
zarar (damage)	kar (profit)	0.30	Antonym
avantaj (advantage)	dezavantaj (disadvantage)	0.30	Antonym
memur (officer)	işçi (employee)	0.44	Not Antonym
aşk (love)	nefret (hate)	0.30	Antonym

6. CONCLUSIONS

In this study, antonym pairs and patterns were automatically extracted from Turkish corpus. Noun-noun and adjective-adjective seeds were created and antonym patterns were generated using these seeds. After generating patterns from initial seeds, reliability score was calculated for each antonym pattern. 11 patterns having reliability score greater than 0.02 were selected to produce new antonym pairs. To measure accuracy of method, noun and adjective target words were used as test words. Using these targets with antonym patterns, candidates were found for each target words. For each antonym pair, antonym score was calculated. Pairs having antonym score greater than 0.3 were assigned to antonym and others were eliminated. For 45 out 196 target words, our method proposed reliable antonym pairs with 77.2% average accuracy.

This study has been shown that Turkish antonym relation patterns can be extracted from corpus easily using some manually created antonym seeds. Candidates also can be easily extracted for a given target word with high accuracy with using reliable antonym patterns. Because patterns are used to extract antonym pairs, high co-occurrence frequency of target with patterns in corpus directly influences success of the method. This is a disadvantage for all of pattern-based methods.

In further studies, we aim to use corpus statistical information with patterns. Thus, antonym pairs occurring with patterns at low frequency can be extracted from corpus.

REFERENCES

[1] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[2] Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, pp. 539-545(1992)

[3] Lobanova, A., van der Kleij, T. and J. Spenader (2010). Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: International Journal of Lexicography. Vol 23: 19-53.

[4] Turney, P.D. (2008), A uniform approach to analogies, synonyms, antonyms, and associations, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp. 905-912.

[5] Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In Proceedings of IJCAI 2003. Acapulco, Mexico.

[6] Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In EMNLP, pages 982–991. Association for Computational Linguistics.

[7] Yazıcı, E. ve Amasyalı, M.F., (2011). Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions, EMO Bilimsel Dergi, İstanbul.

[8] Zeynep Orhan, İlknur Pehlivan , Volkan Uslan , Pınar Önder,Automated Extraction of Semantic Word Relations in Turkish Lexicon, Journal of Mathematical And Computational Applications, 16, 1, Jan. 2011, pp.13-22,

[9] Yildirim, S., Yildiz, T., (2012). “Corpus-Driven Hyponym Acquisition for Turkish Language”, CICLing 13th International Conference on Intelligent Text Processig and Computational Linguistics, 2012.

[10] Şahin, G., Diri, B., Yıldız T., “Pattern and Semantic Similarity Based Automatic Extraction of Hyponym-Hypernym Relation from Turkish Corpus”, 23th Signal Processing and Communications Applications Conference, Malatya, Turkey, (16-19 May), 2015.

[11] Yıldız, T., Yıldırım, S., Diri, B., “Extraction of Part-Whole Relations from Turkish Corpora”, Computational Linguistics and Intelligent Text Processing, CICLing, Greece, 2013.

[12] Yıldız, T., Yıldırım, S., Diri, B., “An Integrated Approach to Automatic Synonym Detection in Turkish Corpus”, 9th International Conference on Natural Language Processing, PolTAL, Springer LNAI proceedings, Warsaw, Poland, (17-19 September), 2014.

[13] Sak, H., Güngör, T. and Saraçlar. M., (2008). “Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus”, The 6th International Conference on Natural Language Processing, GOTAL 2008.

[14] http://zemberek-web.appspot.com/

[15] http://lucene.apache.org/core/

[16] http://tdk.gov.tr/

[17] Lobanova, A., van der Kleij, T. and J. Spenader (2010). Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns. In: International Journal of Lexicography. Vol 23: 19-53.

INTENTIONAL BLANK

TRAVELING SALESMAN PROBLEM IN DISTRIBUTED ENVIRONMENT

Lau Nguyen Dinh and Tran Quoc Chien

University of Da Nang, Danang City, Vietnam
launhi@gmail.com
dhsp@dng.vnn.vn

ABSTRACT

In this paper, we focus on developing parallel algorithms for solving the traveling salesman problem (TSP) based on Nicos Christofides algorithm released in 1976. The parallel algorithm is built in the distributed environment with multi-processors (Master-Slave). The algorithm is installed on the computer cluster system of National University of Education in Hanoi, Vietnam (ccs1.hnue.edu.vn) and uses the library PJ (Parallel Java). The results are evaluated and compared with other works.

KEYWORDS

TSP, processor, parallel, distributed, cycle

1. INTRODUCTION

Traveling salesman problem (TSP) is a well known problem. The problem is solved in different ways. Especially in 1976, Nicos Christofides introduced new algorithms called Christofedes' algorithm [3]. In 2003, Ignatios Vakalis built Christofedes' algorithms on MPI environment [4]. In this paper, we build Christofides' traveling salesman problem in distributed environment. Sequential algorithms are built thoroughly with illustrative examples. In addition, parallel algorithms are experimented in different graphs.

2. CHRISTOFIDES' TRAVELING SALESMAN PROBLEM ALGORITHM

Let G=(V,E) be a graph and let P=V_1, V_2,…, V_k be a path in G. This path is called a Hamiltonian path if and only P is containing every vertex in V. P is a Hamitonian cycle if and only if V_1=V_k and P is a Hamiltonian path. Where G is a directed graph, the terms directed Hamiltonian path and directed Hamiltonian cycle are used. The problem of determining a shortest directed in a weighted directed graph G is called the Traveling Salesman Problem (TSP) [1].

Consider an n x n distance matrix D with positive entries; for example, the distance between the cities the traveling salesman is visiting. We assume D is symmetric, meaning that d_{ij}=d_{ji} for all i and j and d_{ii}=0 for i=1,2,…,n. We claim that [d_{ij}] satisfies the triangle inequality if

$$d_{ij}+d_{jk}\geq d_{ik} \text{ for all } 1\leq i,\ j,\ k \leq n \tag{1}$$

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 19–28, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51503

What the triangle inequality constraint essentially says is that going from city i to city k through city j can not be cheaper than going directed from city i to city k. This is a reasonable assumption, sine the imposed visit to city j appears to be an additional constraint, meaning that can only increase the cost. As a rule of thumb, whenever the entries of the distance matrix represent cost, the triangle inequality is satisfied.

Notice that the graph in this variant of the problem undirected. If we remove any edge from an optimal path for such a graph, we have a spanning tree for the graph. Thus, we can use a algorithm to obtain a minimum spanning tree, then by going twice around the spanning tree, we can convert it to a path that visits every city. Recalling the transformation from Hamiltonian cycle to traveling salesman problem. Christofides [3] introduced a heuristic algorithm based on the minimum spanning tree for this problem.

Definition 2.1. Hamiltonian Cycle is a cycle in an undirected graph that passes through each node exactly once [7].

Definition 2.2. Given an undirected complete weighted graph, TSP is the problem of finding a minimum cost Hamiltonian Cycle [7].

Christofides' Traveling Salesman Problem (Algorithm 1)

Step 1: Find the minimum spanning tree T using the distance matrix D.

Step 2: Find the nodes of T having odd degree and find the shortest complete matching M in the completed graph consisting of these nodes only. Let G' be the graph with nodes {1,2,…,n} and edges in T and M.

Step 3: Find a Hamiltonian cycle in G'.

3.1: Find an Euler cycle C_0=(x,y,z,…,x) in G'.

3.2: Starting at vertex x, we trace C_0 and delete the vertex that has visited before in turn. Then remaining vertices, in the original order in C_0, determine a Hamilton cycle C, which is a required approximation optimal cycle.

The Prim's algorithm can be used in Step 1.

The number of odd-degree nodes in a graph is even. It's easy to see why this is the case: The sum of the degrees of all nodes in a graph is twice the number of edges in the graph, because each edge increases the degree of both its attached nodes by one. Thus, the sum of degrees of all nodes is even. For a sum of integers to be even it must have an even number of odd terms, so we have an even number of odd-degree nodes.

A matching is a subset of a graph's edges that do not share any nodes as endpoints. A perfect matching is a matching containing all the nodes in a graph (a graph may have many perfect matchings). A minimum cost perfect matching is a perfect matching for which the sum of edge weights is minimum. A minimum cost perfect matching of a graph can be found in polynomial time.

Finding a shortest complete matching in a graph is a version of the *minimal weight matching* problem, in which the total weight of the edges obtained from the matching is minimal. Edmonds and Johnson (1970) [5]; William Cook and André Rohe [6] have presented an efficient algorithm for finding minimum weight perfect in any weighted graph.

The Fleury's algorithm [19] can be used in Step 3.1 for finding Euler cycle.

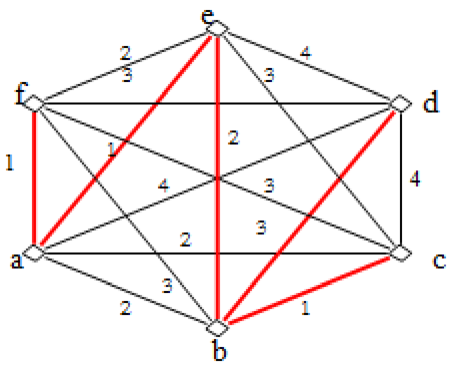


Figure 1. G(V,E) graph

$$D = \begin{Bmatrix} 0 & 2 & 2 & 4 & 1 & 1 \\ 2 & 0 & 1 & 3 & 2 & 3 \\ 2 & 1 & 0 & 4 & 3 & 3 \\ 4 & 3 & 4 & 0 & 4 & 3 \\ 1 & 2 & 3 & 4 & 0 & 2 \\ 1 & 3 & 3 & 3 & 2 & 0 \end{Bmatrix}$$

Figure 2. Distance matrix *D*

Determining whether a graph contains a Hamiltonian cycle is a computationally difficult problem. In fact, the fastest algorithm known has a worst-case time complexity of $O(n^2 2^n)$ in the case of n points. Therefore, the TSP exhibits an exponential worst-case complexity of $O(n^2 2^n)$. Proof [4].

Example: G(V,E) graph is illustrated in Figure 1

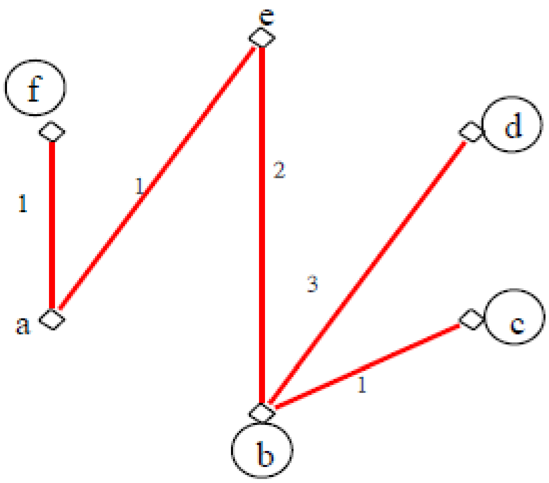


Figure 3. The minimum spanning tree T with the odd-degree vertices encircled

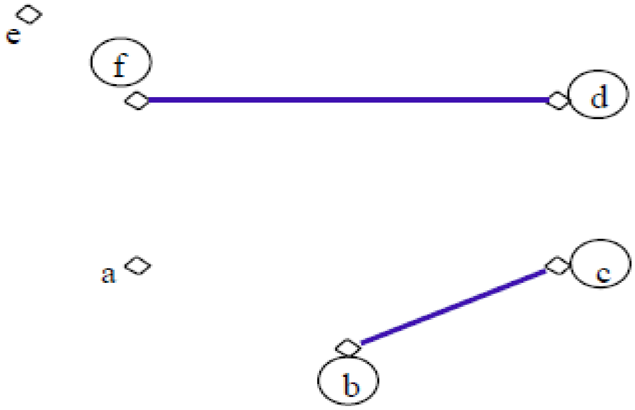


Figure 4. Shows the shortest complete matching M of these odd-degree verteces.

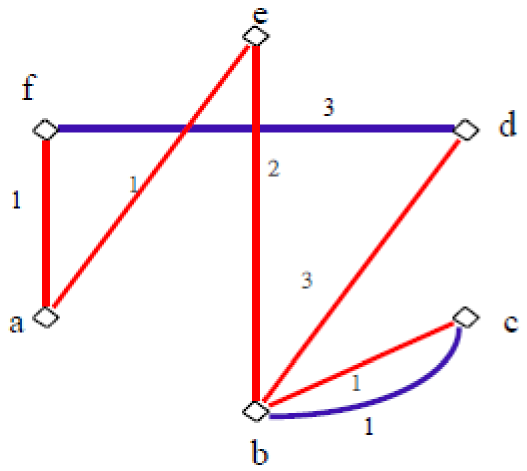


Figure 5. G'(V, E'), E' is edges in T and M

We have Euler cycle C_0 = (a, e, b, c, b, d, f, a). Deleting a repeated vertex b from C_0 results in a Hamilton cycle

C = (a, e, b, c, d, f, a) in G with w(C) = 12. Because the edge (c, d) of C is not in G', C corresponds a salesman route P = (a, e, b, c, b, d, f, a) with w(P) = 12 which visits each vertex of G at least once (Figure 6).

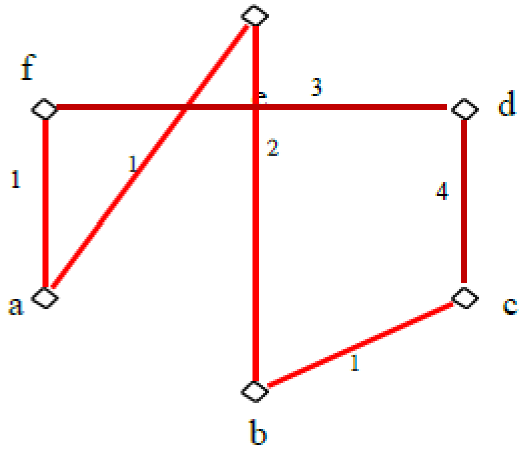


Figure 6. Traveling Salesman tour

For large n, the sequential version of a TSP algorithm becomes impractical. Thus, the need arises to examine a parallel approach in obtaining exact solutions to the TSP problem.

3. The Parallel Traveling Salesman Algorithm

We carry out parallel algorithms on k processors. The parallel is performed in step 1 of the algorithm TSP. Slave processors perform to find MST T. Master processor performs step 2 and step 3 of the algorithm TSP.

Prim's algorithm (Algorithm 2):

```
Input: Let G(V,E) be a graph. V={1,2,…,n}
Output: Minimum Spanning Tree T(B, E') (T be a graph).
Step 1: Initialize T(1,∅). (B={1}, E'={∅})
Step 2: Condition to terminate.
    If T has n-1 edges, then T becomes Minimum Spanning Tree. Otherwise, then go to
    Step 3.
Step 3: (Addition)
    Symbol S is a set as following:
    $S=\{(i, j)\in E | i\in B \text{ and } j\notin B\}$
    Find edge $(u, v)\in S$ so that:
    $d_{uv}=\min\{d_{ij} \mid (i, j)\in S\}$
    If $d_{uv} <\infty$ then v is added to B and (u,v) to E', return to Step 2.  Otherwise, $S=\emptyset$ stop.
```

Parallel Traveling Salesman Problem algorithm (Algorithm 3)

```
Step 1: Create k numbers of Slave processes.
Step 2: Master node send n/k vertex and weight
    matrix D(n x n/k) to Slave.
Step 3: k Slave receives n/k vertex and D(n x n/k)
    from the master node.
Step 4: Master node performs:
    If T(B, E') has n-1 edges, then T becomes Minimum Spanning Tree.
    Otherwise, then go to Step 5.
Step 5: k Slave performs to find:
    - $S_p=\{(i, j)\in E_p | i\in B_p \text{ and } j\notin B_p\}$
    - Find edge $(u, v)_p \in M_p$ so that:
        $d_{uv}^{p} = \min\left\{d_{ij}^{p} \middle| (i,j) \in S_p \ (p=1,2,\ldots k)\right\}$    (2)
    - Send $d_{uv}^{p}$ to Master node.
Step 6: Master node finds $d_{xy}=\min\{ d_{uv}^{p} | \ p=1,2,\ldots k\}$
    If $d_{xy} <\infty$ then y is added to B and (u,v) to E',
    Master node sends y vertex and (x,y) edge to k Slave. return to Step 2.  Otherwise, stop.
Step 7: Master node receives T which is MST, then go to Step 8.
```

Step 8: Master node finds the shortest complete matching M

Step 9: Master node finds the Hamiltonian cycle in G'.

The main loop of the Prim algorithm is executed (*n-1*) times. In each n iteration it scans through all the *m* edges and tests whether the current edge joins a tree with a nontree vertex and whether this is a smallest edge found so far. Thus, the enclosed loop takes time $O(n)$, yielding the worst-case time complexity of the Prim algorithm as $O(n^2)$. Total parallel time $O(n^2/k + n \log k)$.

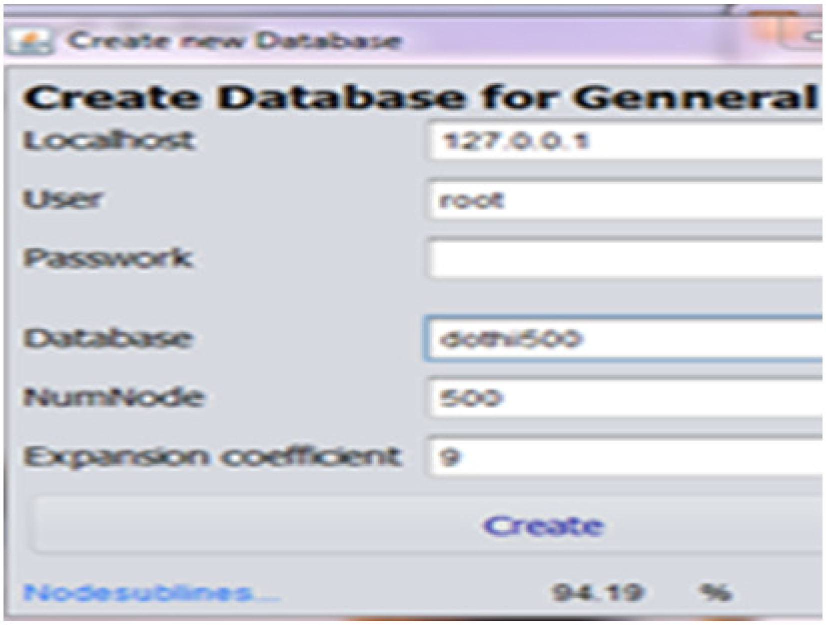
Therefore, algorithm 3 reduces more computation time than algorithm 1.

Parallel computing- Brief Overview:

The development of a wide range of parallel machines with large processing capacities high reliability, and low costs, have brought parallel processing into reality as an efficient way for implementing techniques to solve large scale optimization problems. A good choice of the programming environment is essential to the development of a parallel program.

The processes that are executed on parallel machines, are based on different memory organization methods: shared memory; or distributed memory. In shared memory machines, all processors are able to address the whole memory space. The processors can communicate through operations performed by the parallel tasks on the shared memory. Each task shares a common address space. The advantage of this approach is that the communication can be easy and fast. However, the system is limited by the number of paths between the memory and the processors.

An alternative to the shared memory organization is the distributed memory paradigm. In the framework of the distributed memory organization, the memory is physically distributed among the processors. Each processor can only access its own memory, and communication between processors is performed by messages passed through a communication network. A number of parallel programming tools are available to implement parallel programs for distributed memory environments.



Create new Database

Create Database for Genneral

Localhost 127.0.0.1

User root

Passwork

Database dothi500

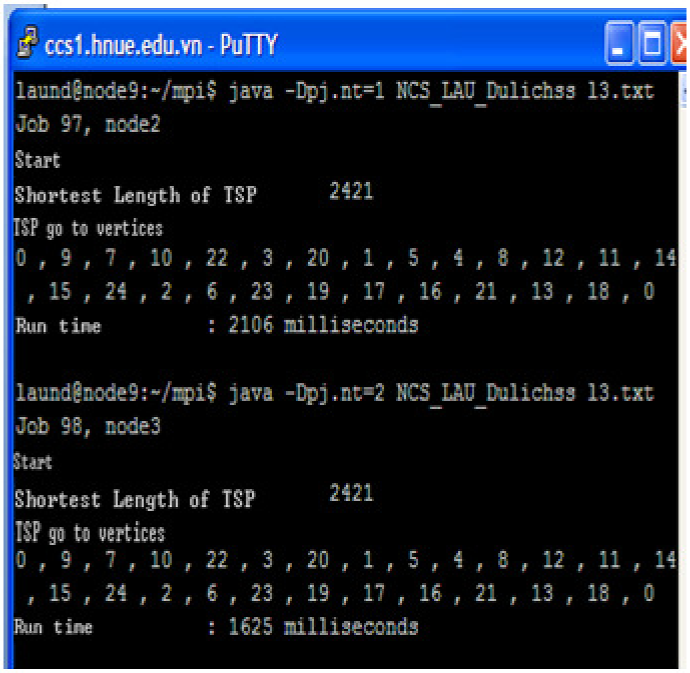
NumNode 500

Expansion coefficient 9

Nodesublines... 94.19 %

Figure 7. Create database (Graph)

So, in this paper we choose the system's computing cluster of Hanoi National University of Education (ccs1.hnue.edu.vn) and use *Parallel java library*_PJ [8], [9].



```
laund@node9:~/mpi$ java -Dpj.nt=1 NCS_LAU_Dulichss 13.txt
Job 97, node2
Start
Shortest Length of TSP      2421
TSP go to vertices
0 , 9 , 7 , 10 , 22 , 3 , 20 , 1 , 5 , 4 , 8 , 12 , 11 , 14
 , 15 , 24 , 2 , 6 , 23 , 19 , 17 , 16 , 21 , 13 , 18 , 0
Run time            : 2106 milliseconds

laund@node9:~/mpi$ java -Dpj.nt=2 NCS_LAU_Dulichss 13.txt
Job 98, node3
Start
Shortest Length of TSP      2421
TSP go to vertices
0 , 9 , 7 , 10 , 22 , 3 , 20 , 1 , 5 , 4 , 8 , 12 , 11 , 14
 , 15 , 24 , 2 , 6 , 23 , 19 , 17 , 16 , 21 , 13 , 18 , 0
Run time            : 1625 milliseconds
```

Figure 8. Parallel Computing Cluster (ccs1.hnue.edu.vn)

Parallel TSP algorithm is built on ccs1.hnue.edu.vn. The program written in Java and use *Parallel java library* (*PJ*). We experimentally sampled nodes as follows: The graph corresponds to 20000 nodes and 30000 nodes. The simulation result is shown in figure 9 and figure 10. This result demonstrates that the runtime of parallel algorithms is better than sequential algorithm.

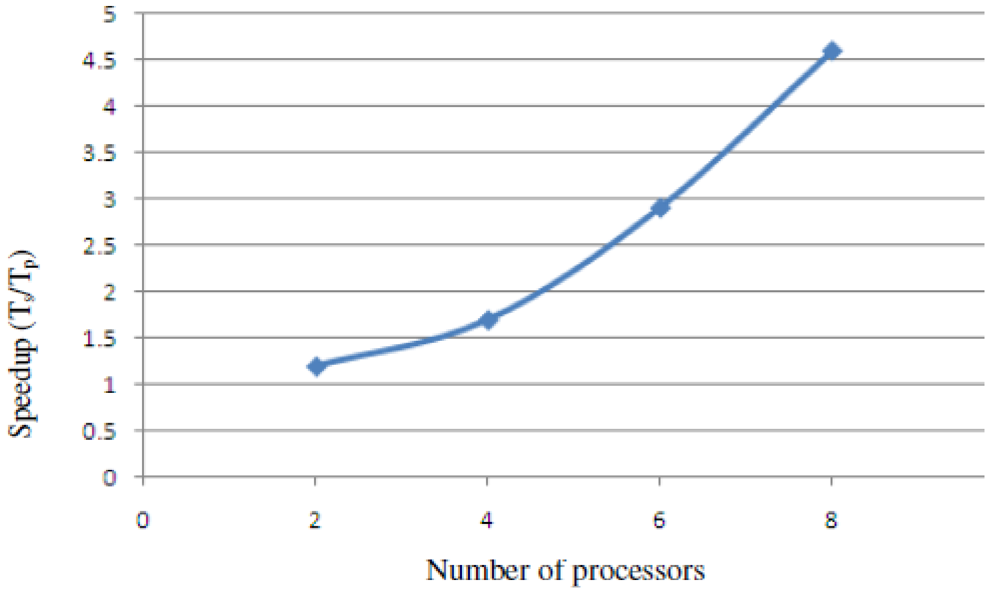


Figure 9. Chart performs the speedup of graph having 20000 nodes

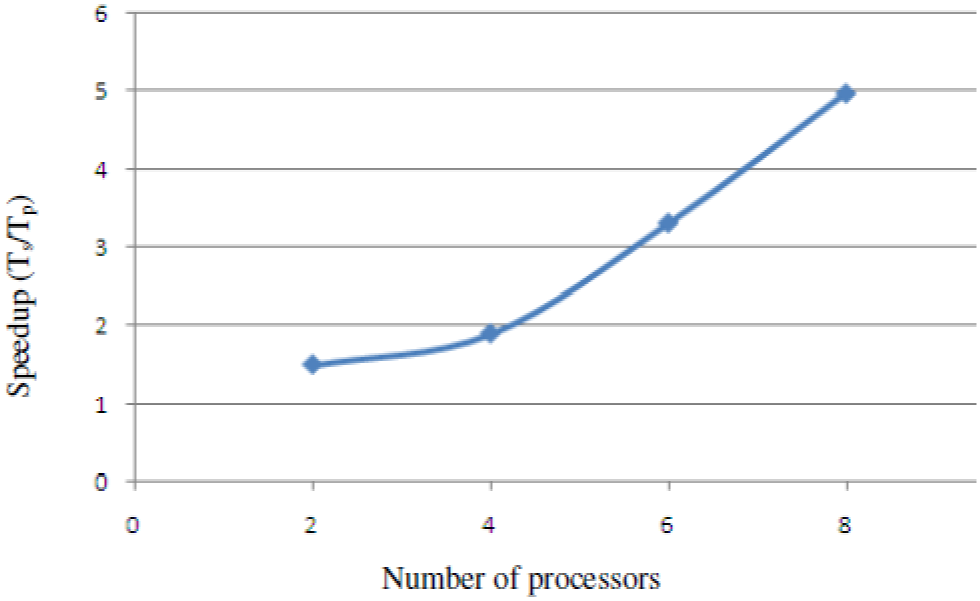


Figure 10. Chart performs the speedup of graph having 30000 nodes

4. CONCLUSION

The detail result of this paper is building sequential and parallel Traveling Salesman Problem algorithm.In addition, to take more advantage of multi-core architecture of the parallel computing system and reduce the computing time of this algorithm, we build this algorithm on multiple processors. Parallel algorithms in this paper are processed at step 1 of the algorithm 1. Ignatios Vakalis 2003 [4] built parallel algorithms by simultaneously looking for DFS (Depth First Search) in step 3 of algorithm 1 to resolve TSP. Random graphs (Figure 7) are created as our database to test the algorithms. As in [4] a small number of vertices graph (less than 12 vertices) are tested. Our algorithms are installed in computer cluster using Parallel Java (PJ) whereas in [4] using MPI. Therefore, our paper has made great contribution to building parallel algorithms using many different libraries.

REFERENCES

[1] Seyed H. Roosta, (1999) Parallel Processing and Parallel Algorithms, Theory and Computation, Springer.

[2] J. Little, K. Murty, D. Sweeney, C. Karel, (1963) “An algorithm for the traveling salesman problem”, Operations Research, No 11 , pp. 972–989.

[3] Nicos Christofides, (1976) Worst-Case Analysis of a New Heuristic for the Travelling Salesman Problem, Carnegie-Mellon University Management Sciences Research Report 388, Pittsburgh, PA.

[4] Ignatios Vakalis, (2003) Traveling Salesperson Problem: A Parallel Approach, NSF Grant No. 9952806, Computational Science Across the Curriculum Project, Capital University, 2003.

[5] J. Edmonds and E.L. Johnson, (1970) "Matching: a well-solved class of integer linear programs", in: Combinatorial structures and their applications (Gordon and Breach, New York, 89-92.

[6] William Cook, André Rohe, (1999) “Computing Minimum-Weight Perfect Matchings”, INFORMS Jounral on Computing, Voll.11, No.2, pp 138-148.

[7] Serge Plotkin, (2010) CS261 - Optimization Paradigms Lecture Notes for 2009-2010 Academic.

[8] Alan Kaminsky. (2007) “Parallel Java: A unified API for shared memory and cluster parallel programming in 100% Java”, 21st IEEE International Parallel and Distributed Processing Symposium (IPDPS 2007), Long Beach, CA, USA.
[9] Jonathan Jude, (2008) “Fast Guide to using the RIT PJ Parallel Java Library: An Introduction to Java Parallel Programming using an API”, ISBN 978-3-8370-2439-5.
[10] Chien Tran Quoc, Lau Nguyen Dinh, Trinh Nguyen Thi Tu, (2013) “Sequential and Parallel Algorithm by Postflow-Pull Methods to Find Maximum Flow”, Proceedings 2013 13th International Conference on Computational Science and Its Applications, ISBN:978-0-7695-5045-9/13 $26.00 © 2013 IEEE, DOI 10.1109/ICCSA.2013.36, published by IEEE- CPS pp 178-181.
[11] Lau Nguyen Dinh, Thanh Le Manh, Chien Tran Quoc, (2013) “Sequential and Parallel Algorithm by Pre-Push Methods to Find Maximum Flow”, Vietnam Academy of Science and Technology AND Posts & Telecommunications Institute of Technology, special issue works Electic, Tel, IT; 51(4A) ISSN: 0866 708X, pp 109-125.
[12] Lau Nguyen Dinh, Chien Tran Quoc and Manh Le Thanh, (2014) “Parallel algorithm to divide optimal linear flow on extended traffic network”, Research, Development and Application on Information & Communication Technology, Ministry of Information & Communication of Vietnam, No 3, V-1.
[13] Lau Nguyen Dinh, Chien Tran Quoc, Thanh Le Manh, (2014) “Improved Computing Performance for Algorithm Finding the Shortest Path in Extended Graph”, proceedings of the 2014 international conference on foundations of computer science (FCS’14), July 21-24, 2014 Las Vegas Nevada, USA, Copyright © 2014 CSREA Press, ISBN: 1-60132-270-4, Printed in the United States of America, pp 14-20.
[14] Chien Tran Quoc, Thanh Le Manh, Lau Nguyen Dinh, (2013) “Sequential and parallel algorithm by combined the push and pull methods to find maximum flow”, Proceeding of national Conference on Fundamental and Applied Infromation Technology Research (FAIR), Hue, Vietnam, 20-21/6/2013.ISBN: 978-604-913-165-3, 538-549.
[15] Chien Tran Quoc, Thanh Le Manh, Lau Nguyen Dinh, (2013) “Parallel algorithm to find maximum flow costlimits on extended traffic network”, Proceeding national Conference XVI "Some selected issues of Information Technology and Communications” Danang 14-15/11/2013, ISBN: 978-604-67-0251-1, 314-321.
[16] Lau Nguyen Dinh, Tran Ngoc Viet, (2012) “Parallelizing algorithm finding the shortest paths of all vertices on computer cluster system”, Proceedings national Conference XVth "Some selected issues of Ìnormation Technology and Communications” Ha Noi, 03-04-2012, 403-409.
[17] Lau Nguyen Dinh, Tran Ngoc Viet, (2012) “A parallel algorithm finding the shortest paths of multiple pairs of source and destination vertices in a graph”, Journal of science and technology - University of DaNang 9 (58), pp. 30-34.
[18] Lau Nguyen Dinh, Tran Ngoc Viet, (2012) “Parallelizing algorithm dijkstra’s finding the shortest paths from a vertex to all vertices”, Journal of science, University of Hue, 74B, 5, pp. 81-92.
[19] Chien Tran Quoc, Graph algorithm: theory and application, 2007

AUTHORS

1. Dr. LAU NGUYEN DINH

Born in 1978 in Dien Ban, Quang Nam, Vietnam. He graduated from Maths_IT faculty of Hue university of science in 2000. He got master of science (IT) at Danang university of technology and hold Ph.D Degree in 2015 at Danang university of technology. His main major: Applicable mathematics in transport, parallel and distributed process, discrete mathemetics, graph theory, grid Computing and distributed programming.

2. Ass. Prof. DrSc. CHIEN TRAN QUOC

Born in 1953 in Dien Ban, Quang Nam, Vietnam. He graduated from Maths_IT faculty. He got Ph.D Degree of maths in 1985 in Charles university of Prague, Czech Republic and hold Doctor of Science in Charles university of Prague, Czech Republic in 1991. He received the tittle of Ass. Pro in 1992. He work for university of Danang, Vietnam. His main major: Maths and computing, applicable mathematics in transport, maximum flow, parallel and distributed process, discrete mathemetics, graph theory, grid Computing, distributed programming.

SEQUENTIAL AND PARALLEL ALGORITHM TO FIND MAXIMUM FLOW ON EXTENDED MIXED NETWORKS BY REVISED POSTFLOW-PULL METHODS

Lau Nguyen Dinh and Tran Quoc Chien

University of Da Nang, Danang City, Vietnam
launhi@gmail.com
dhsp@dng.vnn.vn

ABSTRACT

The problem of finding maximum flow in network graph is extremely interesting and practically applicable in many fields in our daily life, especially in transportation. Therefore, a lot of researchers have been studying this problem in various methods. Especially in 2013, we has developed a new algorithm namely, postflow-pull algorithm to find the maximum flow on traditional networks. In this paper, we revised postflow-push methods to solve this problem of finding maximum flow on extended mixed network. In addition, to take more advantage of multi-core architecture of the parallel computing system, we build this parallel algorithm. This is a completely new method not being announced in the world. The results of this paper are basically systematized and proven. The idea of this algorithm is using multi processors to work in parallel by postflow_push algorithm. Among these processors, there is one main processor managing data, sending data to the sub processors, receiving data from the sub-processors. The sub-processors simultaneously execute their work and send their data to the main processor until the job is finished, the main processor will show the results of the problem.

KEYWORDS

Processor, alogrithm, maximum flow, extended mixed network, parallel.

1. INTRODUCTION

The maximum flow problem on the network is one of the optimization problems on graphs that is widely applicable in practice as well as in combinatorial theory. The problem was proposed and solved by two American mathematicians Ford and Fulkerson in the early 1950 [2] and more and more scientists are interested in research. Edmonds and Karp gave method with complexity $O(|V|.|E|^2)$ [3]. In 1986, A. Goldberg and R.E. Tarjan [4] have developed pre-flow push method with complexity $O(|V|^2.|E|)$ and a lot of paper concerning parallel algorithm are written by many interested researchers [6], [7], [8], [9]. Especially in 2013 we has developed a new algorithm namely, postflow-pull algorithm to

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 29–41, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51504

find the maximum flow on traditional networks [10]. The work of Naveen Garg and Jochen Konemann in 2007 and the above works just concentrate on traditional traffic networks without any specific steps and correct proof. In fact, it is necessary to build extended mixed networks. The problem of finding maximum flow in network mixed network is extremely interesting and practically applicable in many fields in our daily life, especially in transportation. Therefore, a lot of researchers have been studying this problem in various methods. In an ordinary graph the weights of edges and vertexes are considered independently where the length of a path is the sum of weights of the edges and the vertexes on this path. However, in many practical problems, weights at a vertex are not the same for all paths passing this vertex, but depend on coming and leaving edges. The paper develops a model of extended mixed network that can be applied to modelling many practical problems more exactly and effectively. Currently, parallel processing method is a promising and effective solution for the deadlock problems that sequential method encounters such as: program execution time, processing speed, the ability of memory storage, the advantage of multi-core architecture, large-scale data processing. The main contribution of this paper is the revised postflow-push [10] algorithm finding maximal flow on extended mixed network and we build parallel algorithms on multi processors. This is a completely new approach aiming to take advantage of multi-core architecture, to reduce computation time and to solve the problem with large-scale data [10].

2. EXTENDED MIXED NETWORK

Given a graph network G (V, E) with a set of vertices V and a set of edges E, where edges can be directed or undirected, with edge capacity ce:E→R*, so that $ce(e)$ is adge capacity $e \in$ E and vertices capacity cv:V→R*, so that $cv(u)$ is vertices capacity $u \in$ V. [12], [16].

With edge cost be be:E→R*, $be(e)$: cost must be return to transfer an unit transport on edge e.

With each $v\in$ V, Set E_v are set edge of vertice v.

Vertice cost bv:$V\times E_v\times E_v \to R^*$, $bv(u,e,e')$: cost must be return to transfer an unit transport from edge e to vertice u to edge e'.

A set (V, E, ce, cv, be, bv) is called extended mixed network.

3. FLOW EDGE ON EXTENDED MIXED NETWORK

Given an extended mixed network G = (V, E, ce, cv, be, bv). where s is source vertex, t is sink vertex. A set of flows on the edges f = {$f(x,y)$ | $(x,y)\in$ E} is called flow edge on extended mixed network. So that

(i) $0 \le f(x,y) \le ce(x,y)$ $\forall (x,y)\in$ E

(ii) For any vertex k is not a source or sink

$$\sum_{(v,k)\in E} f(v,k) = \sum_{(k,v)\in E} f(k,v)$$

(iii) For any vertex k is not a source or sink

$$\sum_{(v,k)\in E} f(v,k) \leq cv(\mathrm{k})$$

- **Theorem 3.1** Given $f = \{f(x,y) \mid (x,y)\in \mathrm{E}\}$ is flow edge on extended mixed network G, where s is source vertex, t is sink vertex, that is

$$\sum_{(s,v)\in E} f(s,v) - \sum_{(v,s)\in E} f(v,s) = \sum_{(v,t)\in E} f(v,t) - \sum_{(t,v)\in E} f(t,v)$$

Namly total flow go from source vertex equal to total flow going to sink vertex

Proof. $\forall\, x,y\in \mathrm{V} | \nexists\ (\mathrm{x,y}) \in \mathrm{E}$, then assign $f(x,y)= 0$, where

$$\sum_{u\in V}\sum_{v\in V} f(u,v) = \sum_{v\in V}\sum_{u\in V} f(v,u) \Leftrightarrow \sum_{v\in V}\left(\sum_{u\in V} f(u,v) - \sum_{u\in V} f(v,u)\right) = 0$$

$$\Leftrightarrow \sum_{v\in V\setminus\{s,t\}}\left(\sum_{u\in V} f(u,v) - \sum_{u\in V} f(v,u)\right) + \Big(\sum_{(u,s)\in E} f(u,s) - \sum_{(s,u)\in E} f(s,u) \Big) + \Big(\sum_{(u,t)\in E} f(u,t) - \sum_{(t,u)\in E} f(t,u) \Big) = 0$$

From (ii) in section 3, the first term equal to zero, so

$$\Big(\sum_{(u,s)\in E} f(u,s) - \sum_{(s,u)\in E} f(s,u) \Big) + \Big(\sum_{(u,t)\in E} f(u,t) - \sum_{(t,u)\in E} f(t,u) \Big) = 0$$

$$\Leftrightarrow \sum_{(s,u)\in E} f(s,u) - \sum_{(u,s)\in E} f(u,s) = \sum_{(u,t)\in E} f(u,t) - \sum_{(t,u)\in E} f(t,u)$$

The value of flow:

$val(f) = \sum_{(s,u)\in E} f(s,u) - \sum_{(u,s)\in E} f(u,s)$ is called value of flow f.

The maximum problem:

Given an extended mixed network G(V, E, *ce, cv, be, bv*), where s is source vertex, t is sink vertex. The task required by the problem is finding the flow which has a maximum value. The flow value is limited by the total amount of the circulation possibility on the roads starting from source vertex. As a result of this, there could be a confirmation on the following theorem.

- ***Theorem 3.2.*** Given an extended mixed network G(V, E, *ce, cv, be, bv*), where s is source vertex, t is sink vertex , then exist is the maximal flow.

4. MAXIMUM FLOW AND THE MINIMUM CUT

Given an extended mixed network G(V, E, ce, cv, be, bv), where s is source vertex, t is sink vertex. For any set S, T $\subset$V, symbol (S, T) is a set of all edges reached and an unreached going from S input T, (S,T) = {(x, y) $\in$ E |x$\in$ S &y$\in$ T}.

If S, T $\subset$ V| S$\cup$T = V & S$\cap$T = $\varnothing$ and $s\in$ S, $t\in$T, then (S, T) is called cut (source-sink) of G.

Given $f = \{f(x,y) \mid (x,y)\in E\}$ is flow edge on extended mixed network G. Symbols

$$f(\mathrm{S,T}) = \sum_{(x,y)\in(S,T)} f(x,y)$$

• **Theorem 4.1.** Given an extended mixed network G(V, E, ce, cv, be, bv), where s is source vertex, t is sink vertex.

Given $f = \{f(x,y) \mid (x,y)\in E\}$ is flow edge on extended mixed network G and (S, T) is cut of G. Where, $val(f) = f(\mathrm{S,T}) - f(\mathrm{T,S})$

Proof. $\forall$ $x,y\in$ V|$\nexists$ (x,y) $\in$E, then assign $f(x,y)$= 0, we have

$$val(f) = \sum_{(s,u)\in E} f(s,u) - \sum_{(u,s)\in E} f(u,s) = \sum_{u\in V} f(s,u) - \sum_{u\in V} f(u,s) = \sum_{v\in S}\sum_{u\in V} f(v,u) - \sum_{v\in S}\sum_{u\in V} f(u,v)$$

$$(\text{as } \forall v\in \mathrm{S}\backslash\{\mathrm{s}\}, \sum_{u\in V} f(v,u) - \sum_{u\in V} f(u,v) = 0)$$

$$= \sum_{v\in S}\sum_{u\in S} f(v,u) + \sum_{v\in S}\sum_{u\in T} f(v,u) - (\sum_{v\in S}\sum_{u\in S} f(u,v) + \sum_{v\in S}\sum_{u\in T} f(u,v))$$

$$= \sum_{v\in S}\sum_{u\in S} f(v,u) - \sum_{v\in S}\sum_{u\in S} f(u,v) + \sum_{v\in S}\sum_{u\in T} f(v,u) - \sum_{v\in S}\sum_{u\in T} f(u,v)$$

$$= \sum_{v\in S}\sum_{u\in T} f(v,u) - \sum_{v\in S}\sum_{u\in T} f(u,v) = f(\mathrm{S,T}) - f(\mathrm{T,S}).$$

Given (S,T) is cut. Symbol S(T) = {$u\in$ S| $\exists v\in$ T, $(u,v)\in$ (S,T)}

• **Theorem 4.2.** Given an extended mixed network G(V, E, ce, cv, be, bv), where s is source vertex, t is sink vertex.

Given $f = \{f(x,y) \mid (x,y)\in E\}$ is flow edge on extended mixed network G and (S, T) is cut of G. Where, $\forall$S'$\subset$S(T) we have

$$f(\mathrm{S,T}) \le \sum_{v\in S'} c_V(v) + \sum_{(x,y)\in(S,T)\backslash(S',T)} c_E(x,y)$$

Proof. we have

$$f(\mathrm{S,T}) = \sum_{(x,y)\in(S,T)} f(x,y) = \sum_{(x,y)\in(S',T)} f(x,y) + \sum_{(x,y)\in(S,T)\backslash(S',T)} f(x,y) = \sum_{x\in S'}\sum_{(x,y)\in(\{x\},T)} f(x,y) + \sum_{(x,y)\in(S,T)\backslash(S',T)} f(x,y)$$

$$\le \sum_{v\in S'} c_V(v) + \sum_{(x,y)\in(S,T)\backslash(S',T)} c_E(x,y)$$

The capacity of slice cut

Given (S, T) is slice cut of G. Symbol cap(S, T) is capacity of (S, T) slice cut. We have

$$cap(\mathrm{S,T}) = \min\{ \sum_{v\in S'} cv(v) + \sum_{(x,y)\in(S,T)\setminus(S',T)} ce(x,y) \mid \mathrm{S'} \subset \mathrm{S(T)}\}$$

From **Theorem 4.1** and **Theorem 4.2** infered that **Theorem 4.3**

- **Theorem 4.3.** Given $f = \{f(x,y) \mid (x,y)\in \mathrm{E}\}$ is flow edge on extended mixed network G and (S, T) is cut of G. Where $val(f) \le cap(\mathrm{S,T})$.

5. POSTFLOW-PULL METHODS

5.1. Some basic concept

5.1.1. Residual extended mixed network G_f

For flow f on G = (V, E, *ce, cv, be, bv*), where s is source vertex, t is sink vertex. Residual extended network, denoted G_f is defined as the extended mixed network with a set of vertices V and a set of edge E_f with the edge capacity is ce_f and vertices capacity is cv_f as follows:

- For any edge (u, v) ∈ E, if f(u, v)> 0, then (v, u) $\in E_f$ with edge capacity is ce_f (v,u)=f(u, v)

- For any edge (u,v) ∈ E, if c(u,v) -f(u, v)> 0, then (u, v) $\in E_f$ with edge capacity is ce_f(u,v) = ce(u,v) - f(u,v)

- For any vertices v∈ V then $cv_f(v) = cv(v) - \sum_{(x,v)\in E} f(x,v)$.

5.1.2. Preflow

For extended mixed network G = (V, E, *ce, cv, be, bv*). *Preflow* is a set of flows on the edges f = {f(x, y) | (x, y)∈ G} So that

(i) $0 \le f(x,y) \le ce(x,y)\ \forall (x,y) \in E$

(ii) for any vertex k is not a source or sink, inflow is not smaller than outflow, that is

$$\sum_{(v,k)\in E} f(v,k) \ge \sum_{(k,v)\in E} f(k,v)$$

(iii) for any vertex k is not a source or sink

$$\sum_{(v,k)\in E} f(v,k) \le cv(\mathrm{k})$$

5.1.3. Postflow

For extended mixed network G = (V, E, *ce, cv, be, bv*). *Postflow* is a set of flows on the edges f = {f(x, y) | (x, y)∈ G} So that

(i) $0 \le f(x,y) \le ce(x,y)\ \forall (x,y) \in E$

(ii) for any vertex k is not a source or sink, outflow is not smaller than inflow, that is

$$\sum_{(v,k)\in E} f(v,k) \le \sum_{(k,v)\in E} f(k,v)$$

(iii) for any vertex k is not a source or sink

$$\sum_{(v,k)\in E} f(v,k) \le cv(\mathrm{k})$$

Each vertex whose outflow is larger than its inflow is called the unbalanced vertex. The difference between a vertex's inflow and outflow is called excess. The concept of residual extended mixed network G_f is similarly defined as flow.

The idea of this methods is balancing inflow and outflow at the balanced vertices by pushing along an outgoing edge and pushing against an incoming edge. Process of balancing is repeated until no more the unbalanced vertex then we get maximum flow. We store the unbalanced vertices on a generalized queue. A tool called a depth function is used to help select the edge available in residual network to eliminate the unbalanced vertices. Now we assume that a set of the network is denoted as V={0,1,...,|V|-1}.

5.1.4. Depth function

Depth function of the *Postflow* in the extended mixed network G = (V, E, *ce, cv, be, bv*), is a set of non-negative vertex weights d(0), ..., d(|V| −1) such that d(s) = 0(s is source vertex) and d(u)+1 ≥ d(v) for every edge (u,v) in the residual extended mixed network for the flow. An eligible edge is an edge (u,v) in the residual extended mixed network with d(u)+1=d(v).

A trivial depth function is d(0) = d(1) = ... = d(|V| − 1) = 0. Then if we set d(u)= 1, any positive edge to u is the priority edge.

We define a more interesting depth function by assigning to each vertex the latter's shortest–path distance to the sink (its distance to the root in any BFS tree of the network rooted at s. This depth function is valid because d(s)= 0, and for any pair of vertices u and v connected by an edge (u,v) in residual mixed network G_f, then d(u)+1 ≥ d(v), because the path from a to v with edge (u,v) (d(u)+1 must be not shorter than the shortest path from s to v i.e d(v)).

Property 5.1. For any flow f in extended mixed network G and associated depth function d. a vertex's depth d(v) is not larger than the length of the shortest path from vertex s to vertex v in residual extended mixed network G_f .

Proof: For any given vertex v, assume l be the shortest-path length from s to v in the residual extended mixed network G_f. And let (s=v_1, v_2, ..., v_l=v) from s to v. then

$$\begin{aligned} d(v) = d(v_l) &\le d(v_{l-1}) + 1 \\ &\le d(v_{l-2}) + 2 \\ &\ \vdots \\ &\le d(v_1) + l = d(s) + l = l \text{ (because d(s)=0)} \end{aligned}$$

The intuition behind depth function is the following: when an unbalanced node's depth is less then the depth of the sink, it is possible that there is some way to push flow from that node down to the source; else, if an unbalanced node's depth exceeds the depth of the sink, we know that node's flow needs to be pushed back to the sink.

Corollary: if a vertex's depth is greater then |V|, then there is no path from the source to that vertex in the residual extended mixed network G_f.

5.2. General Postflow-pull methods

General Postflow-push methods is briefly described as follows:

Step1:

Initialize: the only Postflow is in the edges leaving for the sink vertices is the following:

$f(\mathrm{v},t)=min\{ce(\mathrm{v},t), cv(v)\}$

The other flows are 0.

Select any available depth function d in the extended mixed network G.

Step 2:

Condition to terminate : If there are no available unbalanced vertices, then postflow f becomes max flow.

Step 3: (pull flow)

Choose unbalanced vertex v.

If exists priority edge (u, v) $\in E_f$ then

If f(v,u)>0, then pull along the edge (u,v) a flow with value $min\{-delta,ce_f(u,v)\}$(where delta<0 is the *excess* of the vertex v).

If $(u,v)\in$ E and $cv_f(u)>0$, then pull along the edge (u,v) a flow with value $min\{-delta, ce_f(u,v), cv_f(u)\}$(where delta<0 is the *excess* of the vertex v).

If it does not exists the priority edge from v, then increased the depth of the vertex v as follows:

$$d(v) := 1 + \min \{d(u) \mid (u, v) \in E_f\}$$

Back to step 2.

◊ Note. In the general Postflow-pull methods, we do not give the detailed steps how to select the initial depth function, how to choose the unbalanced vertices as well as how to choose the priority edges. Performing these detailed steps for many algorithms belongs to the general Postflow-pull methods.

Property 5.2. Postflow-pull methods always preserve the validity of the depth function.

Proof:

(i) Where it exists priority edge (u,v) $\in E_f$: We have d(u)+1 = d (v). After pulling along edge (u,v) a flow, we still have d(v) +1= d(u) +2$\geq$d(u).

(ii) if it does not exist priority edges to v: we have$\forall$u: $(u,v) \in E_f \Rightarrow d(u)+1 > d(v)$. After incrementing d(v):d(v):=1+min{d(u)|(u,v)$\in E_f$} then d(v) still satisfied $\forall$u, (u,v) $\in E_f$: d(u)+1 $\geq$ d(v).

*Property 5.3***.** While Postflow-pull algorithm is in execution, there always exists a directed path from sink vertex to the unbalanced vertex in the residual extended mixed network, and there are no directed paths from source vertex to sink vertex in the residual extended mixed network.

Proof. (by induction)

Initially, the only Postflow is in the edges leaving for the sink vertices is the following: $f(\mathrm{v},t)=min\{ce(\mathrm{v},t),cv(v)\}$ and other flows are 0. Then the first vertices of those edges directed to the sink are unbalanced. With any unbalanced vertex u, we have (t, u) $\in E_f$ and (u,t) $\notin E_f$, inferred there exists paths from t to u, and there are no directed paths from source vertex a to sink vertex in the residual extended mixed network G_f. So the properity is true with the initial flow.

Next, the new unbalanced vertex u only appears when a flow is pushed to the old unbalanced vertex v on the priority edge (u,v). Then the residual extended mixed network will have more edge (v,u). Due to exist of the path from residual extended mixed network from t to v based on inductive hypothesis, there exists a path from t to u in the residual extended network.

To prove that there are no paths from source vertex s to sink t in the residual extended mixed network. It can be argued as follows.

First, vertices u adcajent to sink vertices t, (u, t) $\in$ E, since the initial flow on the edge (u,t) is $f(\mathrm{u},t)=min\{ce(\mathrm{u},t),cv(u)\}$, if (u,t) $\in G_f$, then the flow pushed back along t to u. Where (t, u) is the priority edge, d(t)+1 = d(u) > t(t). Thus each vertex a can reach to t in the residual extended mixed network must have the depth which is greater than the depth of t.

For any u to t in the residual extended mixed network. There exists paths from u to t in the residual extended mixed network: $(u \rightarrow u_1 \rightarrow u_2 \rightarrow \ldots u_k \rightarrow t)$. Similarly argued as above we have $d(u) > d(u_1) > \ldots > d(u_{k-1}) > d(u_k) > d(t)$

Thus each vertex to t must have a depth which is greater than t. Besides, the depth of the source vertex is 0, so it's impossible to reach to t. So there are no directed paths from source vertex to sink vertex in the residual extended network.

• *Corollary*. Vertex's depth is always less than 2.|V|.

Proof. We need to consider only unbalanced vertices, the depth of each unbalanced vertex is either the same as or 1 greater than it was the last time that the vertex was balanced. By the same argument as in the proof of Property 5.1, the path from s source vertex to a given unbalanced vertex in the residual extended mixed network G_f implies that unbalanced vertex's depth is not greater than the sink vertex's depth plus |V| -2 (the source vertex can not be on the path). Since the depth of the sink never changes, and it is initially not greater than |V|, the given unbalanced vertex's depth is not greater than 2.|V| - 2, and no vertex has depth 2|V| or greater.

• Theorem 5.4 General Postflow-pull methods is true.

Proof. First we prove the general Postflow-pull method that terminates after perforing some finite steps. We confirm that after implementing these finite steps there is not any unbalanced vertex. Proof by contradiction method is used. Assume that the set of vertices are infinite, there will exist vertex u that appears infinite times in that set. Since the number of vertices in the network is finite so there exists vertex v≠ u so that the flow is pulled on along (u,v) and (v,u) in infinite times. Since edge (u,v) and edge(v,u) are the priority ones in infinite residual network and d(u)+1= d(v) and d(v)+1=d(u), then the depth of u and v will increment indefinitely, and this conflicts with the above corollary.

When this method terminates, we receive the flow. Based on *property 5.3*, it does not exist a path from the source to the sink in the residual network. According to augmenting-path algorithm, it is max flow.

The complexity of the following method is O(|V|^2|E|) [10].

5.3. Postflow-pull algorithm

This is a particular algorithm in Postflow-pull method. Here the unbalanced vertices are pushed into the queue. With each vertex from the queue, we will pull the flow in the priority edge until the flow becomes either balanced or does not have any priority edge. If it does not exist priority edge but there are unbalanced vertices, then we increase the depth and push it into the queue.

Now we can describe the Postflow-pull algorithm as follows:

Inputs: Extended mixed network G with source s, sink t,

Output: Maximum flow

$$F = (f_{ij}), (\text{i, j}) \in E$$

Step 1: Initialized:

Initialize: the only postflow is in the edges for the source vertices is the following:

$f(\text{v},t) = min\{ce(\text{v},t),\ cv(v)\}$

The other flows are 0.

Choose depth function d(v) which is the length of the shortest path from source s to vertex v.

Push all unbalanced vertices into the queue Q.

Step 2: Condition to terminate: If Q = ∅, then postflow f becomes maximum flow, end.

Step 3:

Get unbalanced vertex v from the queue Q.

Browsing the priority edge (u, v) $\in E_f$

- If f(v,u)>0, then pull along the edge (u,v) a flow with value $min\{-delta, ce_f(u,v)\}$(where delta<0 is the *excess* of the vertex v).

- If $(u,v) \in E$ and $cv_f(u)>0$, then pull along the edge (u,v) a flow with value $min\{-delta, ce_f(u,v), cv_f(u)\}$ (where delta<0 is the *excess* of the vertex v).

- If vertex u is the new unbalanced vertex, then push this vertex u into queue Q.

- If vertex v is still unbalanced, then increased the depth of the vertex v as follows:

$$d(v) := 1 + \min\{d(u) \mid (u, v) \in E_f\}$$

Back to step 2.

6. Postflow-Pull Parallel Algorithm To Find The Maximum Flow

6.1. The idea of the algorithm

Based on the parallel algorithm [10], we build parallel algorithms on m processors. In m processors, there will be a main processor to manage data, divide the set of vertex V of the graph into m-1 sub-processors, and send data to the sub-processors as well as receive data from the sub-processors sending to [6],[7],[8],9], [10].

Sub-processors receive the values from the main processor, then proceed to pull and replace label (pull_relabel) and transfer the results to the main processor.

The main processor after receiving the results from the sub-processors will perform replacement label (Relabel) until finding the maximum flow

6.2. Building the parallel algorithm

Inputs: Extended mixed network G with source s, sink t m processors (P_0, P_1,…, P_{m-1}), where P_0 is the main processor

Output: Maximum flow

$$F = (f_{ij}), (\mathrm{i, j}) \in E$$

Step 1: The main processor P_0 performs

(1.1). initialize: e, d, f, c_f, Q: set of unbalanced vertices (excluding the vertices s and t) are the vertices with positive excess.

(1.2). divide set of vertices V into sub-processors:

Let P_i be the i^{th} sub-processor (i = 1,2, ..., m-1)

P_i will receive the set of vertices V_i so that

$$(V_i \cap V_j = \phi \text{ if } i \neq j, \text{ and } \cup_i \{V_i\} = V)$$

(1.3). The main processor sends e, c_f to sub-processors

Step 2: The Condition to terminate: If Q = ∅, then postflow f becomes maximum flow, end. Else, go to step 3.

Step 3: The main processor sends d to sub-processors

Step 4: m-1 sub-processors (P_1, P_2, …,P_{m-1}) implement

(4.1) Receive e, c_f, d and the set of vertices from the main processor

(4.2) Handling unbalanced vertice v (pull and replace label). Get unbalanced vertexs v from Q and $v \in V_i$ (i= 1,2, ..., m-1). Browsing the priority edge (u, v) $\in E_f$

- If f(v,u)>0, then pull along the edge (u,v) a flow with value *min*{*-delta*,$ce_f(u,v)$}(where delta<0 is the *excess* of the vertex v).

- If $(u,v) \in E$ and cv_f(u)>0, then pull along the edge (u,v) a flow with value *min*{*-delta*, $ce_f(u,v)$, $cv_f(u)$}. (where delta<0 is the *excess* of the vertex v).

If vertex v is still unbalanced, then increased the depth of the vertex v as follows:

$$d(v) := 1 + \min \{d(u) \mid (u, v) \in E_f\}$$

(4.3) Send e, c_f, d to the main processor

Step 5: The main processor implements

(5.1) Receive e, c_f, d from step 4.3

(5.2) This step is distinctive from the sequential algorithms to synchronize our data, after receiving the data in (5.1), the main processor checks if all the edges $(u,v)\in E$ that have d(v)> d(u)+1, the main processor will relabel for vertices u, v as follows:

- e(u):= e(u)-ce$_f$(u,v), e(v):= e(v)+ce$_f$(u,v)

- If f(v,u)>0, then f(u,v):= *min*{*-delta*,*ce$_f$*(*u,v*)} (where delta<0 is the *excess* of the vertex v).

- If (*u,v*)∈E and *cv$_f$*(u)>0, then f(u,v):= *min*{*-delta*, *ce$_f$*(*u,v*), *cv$_f$*(u)} (where delta<0 is the *excess* of the vertex v). Put the new unbalanced vertex into set Q

(5.3) If $\forall u\in V$ e(u)=0, eliminate u from active set Q. Back to step 2.

Theorem 3.1. Postflow-pull parallel algorithm is true and has complexity O(|V|^2 |E|).

Proof: Similar to [10]. postflow-pull parallel algorithm is build in accordance with other parallel computing system such as: PRAM, Cluster system, CUDA, RMI, threads,… Push and replace label using *atomic*, due to support of *atomic 'read-modify-write'* instructions, are executed atomically by the architecture. Other than the two execution characteristics provided by the architecture, we do not impose any order in which executions from multiple sub-processors can or should be interleaved, as it will be left for the sequential consistency property of the architecture to decide.

The outcome of the execution reduces to only a few simplified scenarios. By analyzing these scenarios, we can show that function f is maintained as a valid depth function. A valid d guarantees that there does not exist any paths from s to t throughout the execution of the algorithm, and hence guarantees the optimality of the final solution if the algorithm terminates. The termination of the algorithm is also guaranteed by the validatity of d, as it bounds the number of pull and relabel operations to O(|V |^2|E|). □

Parallel algorithm for finding maximum flow in the extended mixed network is built on m processors. The program written in Java with database administration system MySQL We experimentally sampled nodes as follows: The extended mixed graph corresponds to 18000 nodes and 25000 edge. The simulation result is shown in figure 1. This result demonstrates that the runtime of parallel algorithms is better than sequential algorithm.

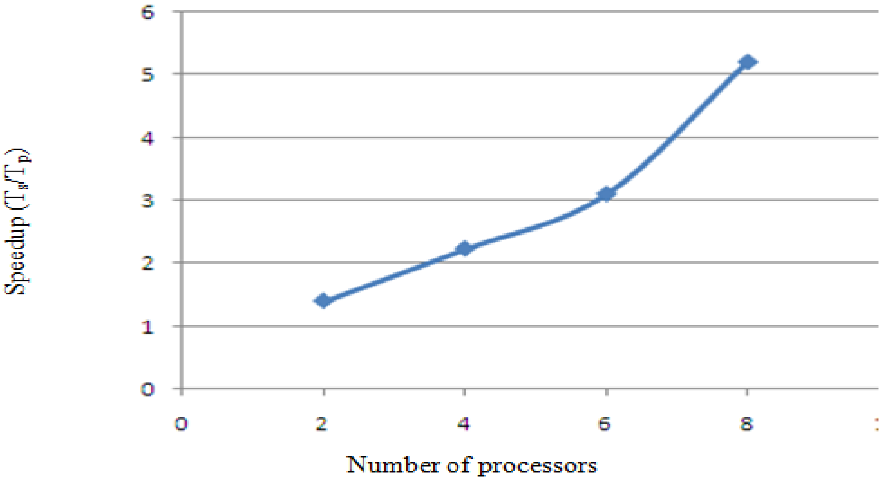


Figure 1. Chart performs the speedup of extended Mixed graph having 18000 nodes and 25000 edge

7. CONCLUSION

The detail result of this paper is building sequential and parallel algorithm by postflow-pull methods to find maximum flow in extended mixed network. In addition, to take more advantage of multi-core architecture of the parallel computing system and reduce the computing time of this algorithm, we build this algorithm on multiple processors. This is a completely new method not being announced in Vietnam and in the world. The results of this paper are basically systematized and proven.

REFERENCES

[1] Chien Tran Quoc, 2010 “Postflow-pull methods to find maximal flow”, Journal of science and technology - University of DaNang, 5(40), pp 31-38.

[2] L. R. Ford and D. R. Fulkerson, 1962 Flows in Networks Princeton University Press.

[3] J. Edmonds and R. M. Karp, 1972 “Theoretical improvements in algorithmic efficiency for network flow problems,” J. ACM, vol. 19, no. 2, pp. 248–264.

[4] A. V. Goldberg and R. E. Tarjan, 1986 "A new approach to the maximum flow problem," in STOC '86: Proceedings of the eighteenth annual ACM symposium on Theory of computing. New York, NY, USA: ACM, pp. 136-146.

[5] Robert Sedgewick, 2011 Algorithms in C, Part 5: Graph Algorithms. Addison-Wesley.

[6] R. J. Anderson and a. C. S. Jo, 1992 “On the parallel implementation of goldberg’s maximum flow algorithm” in SPAA ’92: Proceedings of the fourth annual ACM symposium on Parallel algorithms and architectures. New York, NY, USA: ACM, pp. 168–177.

[7] D. Bader and V. Sachdeva, 2005 “A cache-aware parallel implementation of the push- relabel network flow algorithm and experimental evaluation of the gap relabeling heuristic” in PDCS ’05: Proceedings of the 18th ISCA International Conference on Parallel and Distributed Computing Systems.

[8] B. Hong, 2008 “A lock-free multi-threaded algorithm for the maximum flow problem” in IEEE International Parallel and Distributed Processing Symposium, Aprail.

[9] Zhengyu He, Bo Hong, 2010 “Dynamically Tuned Push-Relabel Algorithm for the Maximum Flow Problem on CPU-GPU-Hybrid Platforms”, School of Electrical and Computer Engineering-Georgia Institute of Technology.

[10] Chien Tran Quoc, Lau Nguyen Dinh, Trinh Nguyen Thi Tu, 2013 “Sequential and Parallel Algorithm by Postflow-Pull Methods to Find Maximum Flow”, Proceedings 2013 13th International Conference on Computational Science and Its Applications, ISBN:978-0-7695-5045-9/13 $26.00 © 2013 IEEE, DOI 10.1109/ICCSA.2013.36, published by IEEE- CPS pp 178-181.

[11] Lau Nguyen Dinh, Thanh Le Manh, Chien Tran Quoc, 2013 “Sequential and Parallel Algorithm by Pre-Push Methods to Find Maximum Flow”, Vietnam Academy of Science and Technology AND Posts & Telecommunications Institute of Technology, special issue works Electic, Tel, IT; 51(4A) ISSN: 0866 708X, pp 109-125.

[12] Lau Nguyen Dinh, Chien Tran Quoc and Manh Le Thanh, 2014 “Parallel algorithm to divide optimal linear flow on extended traffic network”, Research, Development and Application on Information & Communication Technology, Ministry of Information & Communication of Vietnam, No 3, V-1.

[13] Naveen Garg, Jochen Könemann, 2007 “Faster and Simpler Algorithms for Multicommodity Flow and Other Fractional Packing Problems”, SIAM J. Comput, Canada, 37(2), pp. 630-652.

[14] Lau Nguyen Dinh, Chien Tran Quoc, Thanh Le Manh, 2014 “Improved Computing Performance for Algorithm Finding the Shortest Path in Extended Graph”, proceedings of the 2014 international conference on foundations of computer science (FCS’14), July 21-24, 2014 Las Vegas Nevada, USA, Copyright © 2014 CSREA Press, ISBN: 1-60132-270-4, Printed in the United States of America, pp 14-20.

[15] Chien Tran Quoc, Thanh Le Manh, Lau Nguyen Dinh, 2013 “Sequential and parallel algorithm by combined the push and pull methods to find maximum flow”, Proceeding of national Conference on

Fundamental and Applied Infromation Technology Research (FAIR), Hue, Vietnam, 20-21/6/2013.ISBN: 978-604-913-165-3, pp 538-549.

[16] Chien Tran Quoc, Thanh Le Manh, Lau Nguyen Dinh, 2013 “Parallel algorithm to find maximum flowcostlimits on extended traffic network”, Proceeding national Conference XVI "Some selected issues of Information Technology and Communications” Danang 14-15/11/2013, ISBN: 978-604-67-0251-1, pp 314-321.

[17] Lau Nguyen Dinh, Tran Ngoc Viet, 2012 “Parallelizing algorithm finding the shortest paths of all vertices on computer cluster system”, Proceedings national Conference XVth "Some selected issues of Ìnormation Technology and Communications” Ha Noi, 03-04-2012, pp 403-409.

[18] Lau Nguyen Dinh, Tran Ngoc Viet, 2012 “A parallel algorithm finding the shortest paths of multiple pairs of source and destination vertices in a graph”, Journal of science and technology - University of DaNang 9 (58),2012, pp 30-34.

[19] Lau Nguyen Dinh, Tran Ngoc Viet, 2012 “Parallelizing algorithm dijkstra’s finding the shortest paths from a vertex to all vertices”, Journal of science, University of Hue, 74B, 5, pp 81-92.

AUTHORS

1.Dr. LAU NGUYEN DINH

Born in 1978 in Dien Ban, Quang Nam, Vietnam. He graduated from Maths_IT faculty of Hue university of science in 2000. He got master of science (IT) at Danang university of technology and hold Ph.D Degree in 2015 at Danang university of technology. His main major: Applicable mathematics in transport, parallel and distributed process, discrete mathemetics, graph theory, grid Computing and distributed programming.

2.Ass. Prof. DrSc. CHIEN TRAN QUOC

Born in 1953 in Dien Ban, Quang Nam, Vietnam. He graduated from Maths_IT faculty. He got Ph.D Degree of maths in 1985 in Charles university of Prague, Czech Republic and hold Doctor of Science in Charles university of Prague, Czech Republic in 1991. He received the tittle of Ass. Pro in 1992. He work for university of Danang, Vietnam. His main major: Maths and computing, applicable mathematics in transport, maximum flow, parallel and distributed process, discrete mathemetics, graph theory, grid Computing, distributed programming.

INTENTIONAL BLANK

INTRODUCING SIMPLEX MASS BALANCING METHOD FOR MULTI-COMMODITY FLOW NETWORK WITH A SEPARATOR

Ziauddin Ursani[1,2], Ahsan A. Ursani[3,4] and David W. Corne[1,2]

[1]Heriot Watt University UK
z.ursani@hw.ac.uk
d.w.corne@hw.ac.uk
[2]Route Monkey Ltd.
[3]South Asian University New Delhi India
[4]Mehran University of Engineering and Technology Sindh Pakistan
ahsan.ursani@faculty.muet.edu.pk

ABSTRACT

Maximization of flow through the network has remained core issue in the field of optimization. In this paper a new advanced network problem is introduced, i.e., a multi-commodity network with a separator. This network consists of several sub-networks linked through a commodity separator. The objective is to maximize the flow of the commodity of interest from the commodity mixture at the output of the commodity separator, while observing the capacity constraints of all sub-networks. Such networks are present in Oil and Gas development fields. Such networks are also conceptual representation of material flows of many other manufacturing systems. In this paper an algorithm is developed for maximization of flow in these networks. Maximization of flow in such networks has direct practical relevance and industrial application. The developed algorithm brings together two distinct branches of computer science i.e., graph theory and linear programming to solve the problem.

KEYWORDS

Graph Theory, Linear Programming, Multi-commodity Network Flow Optimization, Commodity of Interest, Hybrid Algorithm

1. INTRODUCTION

Flow maximization through networks has been a major problem under study for last several decades [1]. This is because many real world problems can be formulated as a network problem such as optical networks [2], wireless networks [3], reliability networks [4, 5], biological networks [6], production assembly networks [7] and social networks [8] etc.

A typical flow network is a directed graph with number of nodes connected through number of edges. Each edge has a limited capacity. A network also has a source node and a sink node. It is assumed that source node can produce flow of unlimited capacity. The problem is to push maximum flow through the network from the source node to the sink node such that capacity of

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 43–57, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51505

any edge is not violated and all nodes must be balanced nodes i.e. flow going into the node is equal to flow going out of the node.

In 1956 a remarkable theorem on this network was developed which is popularly known as max-flow-min-cut theorem [9, 10]. According to this theorem maximum flow through the network is equal to minimum cut of the network. The minimum cut of the network is defined as a cut of minimum size through the network that disconnects completely the source from the sink such that no flow from the source could pass to the sink. If this cut consists of only single point then any bottleneck in this cut can cause it to become single point of failure (SPOF) triggering failure of the entire system [11]. This theorem might not be applicable in some networks with multiple commodities coming from multiple sources and going into multiple sinks [12], however this theorem provides strong base to construct max flow theorems for many of these types of networks [13]. Based on this theorem a number of approaches have been discovered to solve this problem. These approaches can be divided into two main branches, i.e. augmentation paths algorithms, [9, 10, 14, 15, 16, 17, 18, 19] and pre-flow push algorithms [20, 21, 22, 23, 24, 25, 26, 27, 28]. Some novel ideas have also been discovered such as pseudo flows [29], and draining algorithm [30].

There is also an advanced network problem called multi-commodity flow network problem. A multi-commodity network is the network carrying mixture of commodities from multiple sources. The sources are considered to be of unlimited capacity, each producing a mixture of commodities with different proportions. The problem is to maximize the flow of the commodity of interest (COI) through the network such that final mixture coming out through the sink has the maximum proportion of the COI. This multi-commodity network problem is closely related to oil and gas development field where there is a number of wells connected to a network. Each well produces mixture of oil, gas, and water in different proportions. Industry usually wants to maximize the flow of oil through the network as oil is considered the most precious commodity. A quick solution to this problem was presented in mass balancing theorem [31].

However above problem presents only half of the picture of real world. In the real world industrial scenario, there is a multi-commodity network that terminates onto a separator that separates all these commodities, each of which flows through its own flow network towards its respective sink. Faults occur regularly in these huge networks, which directly affect capacity of the relevant sub-network requiring readjustment of production from source to maximize the output of the COI. Therefore, capacity of an individual commodity network has direct effect on the production system of the whole mixture of commodities. This problem is also conceptual representation of material flows of many other manufacturing systems such as mining Industry. The mining industry has to deal with a raw material coming from different sources. The industry doesn’t have much control over contents of this raw material as each source produces raw material of its own configuration i.e. its own mix of compounds in different proportions. The industry has to process and refine this raw material for further use. However industry has limited processing and refining capacity due to industrial, operational, qualitative and environmental reasons [32]. Therefore the industry needs to determine production rates from different sources in such a way that requirements of its processing unit regarding volume and predefined content of incoming material are met. A similar problem was solved by iterative mass balance method [33] but in that method capacity constraints of network were not considered therefore the method fails short of real world scenario.

In this paper network with a commodity separator problem is formally introduced and its solution is also proposed. The proposed method combines two distinct fields of Computer Science i.e. linear programming [34] and graph theory [35]. Graph algorithms are usually considered faster and simpler than linear programming in the area of flow networks. This is because each node of the network adds one equation to the set of equations to be solved under linear programming. Therefore linear programming becomes very cumbersome with very large networks. On the other hand, graph theory does not have mathematical apparatus to deal with separation of commodities in the network. Therefore, the proposed method makes use of both linear programming and graph theory to achieve easy and quick solution.

The scenario of the network as presented above represents dynamic situation where faults occur and get repaired very frequently. For such a situation a quick method of optimization based on Mass Balancing Theorem was introduced. In mass balancing theorem an interesting property of network was discovered that the fully saturated network is actually balance of certain easily computable flow load on the either side of the minimum cut. Utilizing this property a flow dissipation algorithm was developed to maximize the flow through the network. An interesting thing about this algorithm is that it visits only unbalanced nodes rather than the whole network to maximize the flow. Therefore this algorithm has very important role to play in dynamic networks where the network continues changing its state. A change is marked by removal of E^- edges and/or addition of E^+ edges. Due to these changes each time certain number of nodes becomes unbalanced. The upper bound on this number δ is given by:

$$\delta = 2(E^+ + E^-) \tag{1}$$

The number in equation 1 is only a small fraction of total number of nodes in the network and by visiting only those unbalanced nodes flow can be maximized. Furthermore this algorithm is also extended to the multi-commodity network, where a COI was maximized in presence of multiple sources. However, as in this paper advanced problem of multi-commodity network with a separator is considered. For this advanced problem as discussed earlier the method only based on graph theory is not enough. In such problems flow cannot be maximized without add of linear programming to deal with the separation of flows into individual commodities. Keeping above situation in mind a method has been devised that is hybrid of simplex method based on linear programming and mass balancing method based on graph theory for the particular problem formulated in this paper. However roles of both mass balancing theorem and simplex method are chosen in a way to utilize positive points of both the methods. Flow through all the sub-networks is maximized using mass balancing theorem while simplex method is applied only on a commodity separator to optimize the flow of COI. The hybrid algorithm is called simplex mass balancing (SMB) Method.

The remainder of the paper is organized as follows. Section 2 presents mathematical formulation of the problem, section 3 explains the SMB method, section 4 provides the proof of optimality, in section 5, a solved example of proposed method is presented. Section 6 analyses the complexity of the proposed algorithm, and finally section 7 concludes the paper and discusses the future work.

2. PROBLEM FORMULATION

Let the multi-commodity network consists of n sources and m commodities, and each source j produces a unique mix of commodities in quantity Q^jsuch that

$$\underset{j=1,n}{\forall} Q^j = \sum_{i=1}^{i=m} q_i^j = \sum_{i=1}^{i=m} \gamma_i^j Q^j \tag{2}$$

q_i^j= flow of commodity *i* in source *j*

γ_i^j = proportion of flow of commodity *i* in source *j* such that

$$0 \leq \gamma_i^j \leq 1 \tag{3}$$

Equation 2 shows that total quantity of mixture is sum of all the quantities of individual commodities in the mixture, where quantity of each individual commodity can be determined from its proportion in the mixture. The value of proportion varies between 0 and 1 (expression 3).

Flow from all the sources ultimately terminate onto a separator, where the commodity mixes are separated. At the output of the separator, there are *m* commodity networks each corresponding to a single commodity, carrying *i*th commodity to the *i*th sink. The goal is to maximize the output of commodity of interest (COI) while obeying the capacity constraints of multi-commodity network and each of the *m* commodity networks.

Figure 1 shows a multi-commodity network, namely, N_0 connected to *n* sources S_1...... S_n and a separator *U*. In addition, there are *m* commodity networks, namely N_1.....N_m, which originate from the separator *U* and each of these networks has its own sink, i.e. T_1 through T_m, respectively. For the problem formulation, following subsections define some notions.

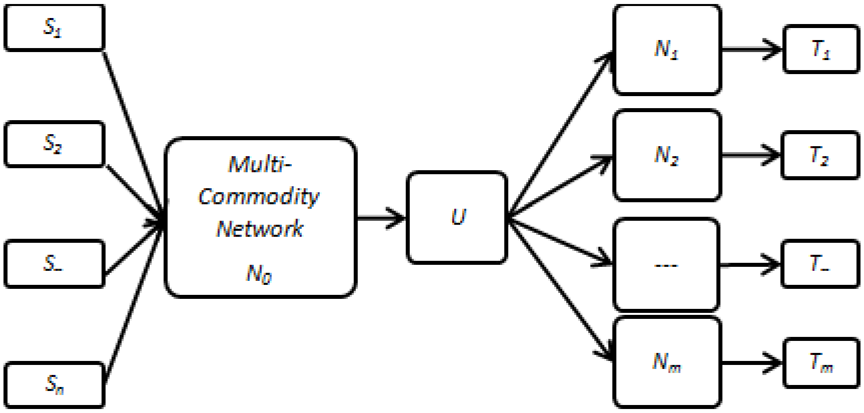


Figure 1. Multi-Commodity Network with Separator

2.1. Unified and Individual Source Networks (USN and ISNs)

Let us modify the network in Figure 1 by connecting its source nodes S_1...... S_n with the universal source node S_0 of unlimited capacity through the edges E_1, E_2, ..., E_n of unlimited capacity respectively. Furthermore considering the separator as the ultimate sink, the network of Figure 1 can be reduced to the network shown in Figure 2. The network hereby referred to as Unified Source Network (USN). The USN in Figure 2 can be calibrated into the individual source networks. The individual source network corresponding to the source i, ISN_i is the network with capacity of $E_i = \infty$ and capacity of $E_j = 0$ where $j \in \{1, \dots.., n/j \neq i\}$. This means that in the individual network of source i all the other sources will be disconnected from the network except source i itself. Furthermore the capacity of source i is also considered unlimited.

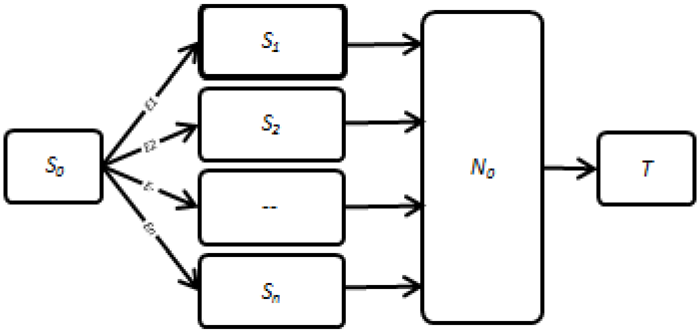


Figure 2. Unified Source Network and n Individual Source Networks

2.2. Individual Commodity Network (ICN)

Considering the separator U as the primary source for each commodity network, the network of Figure 1 can be reduced to the network shown in Figure 3. In Figure 3 since there are m commodities hence there are m ICNs, such that for ICN$_i$, capacity of $e_i = \infty$ and capacity of $e_j = 0$ where $j \in \{1, \ldots\ldots, n/j \neq i\}$. This means that in the individual commodity network of commodity i all the other commodities are disconnected from the network except commodity i itself. Furthermore the capacity of primary source of commodity i is also considered unlimited.

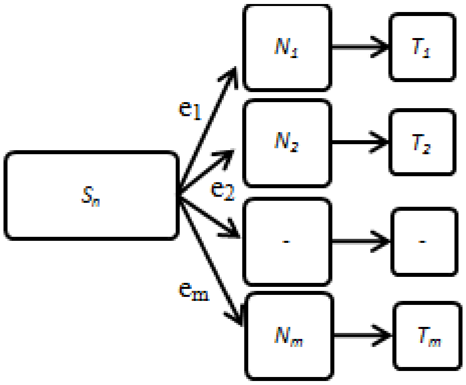


Figure 3. m Individual Commodity Networks

Let us consider

C_0 = Minimum cut of the USN in Figure 2
$C_{si\in\{1,\ldots,n\}}$= Minimum cut of the $ISN_{i\in\{1,\ldots,n\}}$ respectively in Figure 2
$C_{ci\in\{1,\ldots,m\}}$ = Minimum cut of the $ICN_{i\in\{1,\ldots,m\}}$ respectively in Figure 3

Therefore the maximum flow Q_0 in the multi-commodity network of Figure 1 is given by

$$Q_0 \leq min\left(C_0, \sum_{i=1}^{i=n} C_{si}, \sum_{i=1}^{i=m} C_{ci}\right) \tag{4}$$

This means that maximum flow in the network can be only be minimum of the following three quantities.

1. Minimum cut of the unified source network
2. Sum of minimum cuts of individual source networks
3. Sum of minimum cuts of individual commodity networks

Since in any case

$$C_0 \leq \sum_{i=1}^{i=n} C_{si} \tag{5}$$

Therefore expression 4 reduces to

$$Q_0 \leq min\left(C_0, \sum_{i=1}^{i=n} C_{ci}\right) \tag{6}$$

Expression 6 shows that maximum flow through the multi-commodity network of Figure 1 cannot be greater than lesser of the two quantities i.e., minimum cut of USN of Figure 2, and sum of minimum cuts of all the individual commodity networks ICNs of Figure 3. The ≤ sign in this expression indicates that there are other constraints too that may restrict the flow. Those constraints are shown in expressions 7 and 8. Suppose q_i^j is flow of commodity *i* in source *j* then

$$\forall_{j=1,n}\, Q^j \leq C_{sj} \tag{7}$$

and $$\forall_{i=1,m}\, Q_i = \sum_{j=1}^{j=n} q_i^j \leq C_{ci} \tag{8}$$

Expression 8 shows that flow from any source *j* must not be greater than minimum cut of its individual source network and expression 7 shows that total flow of any commodity *i* must not be greater than the minimum cut C_{ci} of its respective individual commodity network. The ≤ sign signifies the fact that if one of the commodities k exhausts the capacity of its individual network C_{ck}, then flow cannot be further increased for other commodities $i \in \{1, \ldots, m/i \neq k\}$ as increase in the overall mixture would also increase the flow of the commodity *k*. Therefore objective is to maximize the flow of COI, Q_{coi} i.e.,

$$max(Q_{coi}) = max\left(\sum_{j=1}^{j=n} q_{coi}^j\right) \tag{9}$$

Substituting the values of q_{coi}^j from expression 2 into expression 9 gives the following linear function

$$Z = \sum_{j=1}^{j=n} \gamma_{coi}^j Q^j \tag{10}$$

The linear function in equation 10 is to be maximized under the constraints of expressions 6 through 8.

3. SMB METHOD FOR MULTI-COMMODITY NETWORK WITH A SEPARATOR

A method for maximization of flow of a commodity of interest through the network with a separator has been devised by keeping problem formulation presented in section 2 in mind. The method hybridizes two distinct fields of Computer Science i.e., linear programming and graph theory as explained in section 1. Linear programming is used to maximize linear function shown in equation 10 under the constraints in expressions 6-8. However to determine the value of constraints mass balancing method of flow maximization is chosen. The reasons for choosing this method has already been discussed in section 1. This hybridized method is termed as simplex mass balancing (SMB) method.

The algorithm is explained in the following steps.

1. Create all the networks including USNs, ISNs (Section 2.1) and ICNs (Section 2.2)
2. Maximize the flow through each USNs, ICNs and ISNs to determine the values of x_0, $x_{i\in\{1,....,n\}}$, $y_{i\in\{1,....,m\}}$ respectively, corresponding to n sources and m commodities.
3. Design linear programming formulation (equation 10) under the constraints 6-8 from output of step 2.
4. Maximize the linear function of equation 10 using Simplex method to determine $Q^{j\in\{1,...,n\}}$.
5. Maximize the flow through the USN by equating capacity of $E_{j\in\{1,2,....,n\}}$ with $Q^{j\in\{1,...,n\}}$ respectively.
6. Compute quantity of each commodity q_i using equation 10 from output of step 4.
7. For all i, maximize the flow in ICN_i by equating capacity of $e_{j\in\{1,2,....,m\}}$ with $q_{j\in\{1,...,m\}}$ respectively.
8. Join USN obtained from step 5 and ICNs obtained from step 7 and remove additional edges and universal source node to represent actual network with maximized flow of COI.

In steps 2, 5 and 7 flow is maximized using mass balancing theorem. From the above procedure it can be seen that in in the second step a method based on graph theory is used to determine minimum cuts of various conceptual networks introduced in section 2. The values of those minimum cuts are later used in design of linear programming formulation in step 3. The designed linear function is then optimized in step 4 using simplex method to determine optimal flows from all sources. In step 5, a flow through multi-commodity network is again maximized by restricting flow from sources to optimal flows obtained in previous step. In step 6, quantity of each commodity is computed in the resultant output mixture from all the sources. In step 7, flow in each commodity network is maximized by restricting commodity quantity obtained in previous step. In the final step, all the conceptual networks are joined together to form original network.

4. PROOF OF OPTIMALITY

Simplex method is a well-known method which optimizes linear function [34]. SMB method uses this method to maximize flow of commodity of interest from the mixture of commodities. If the LP formulation is correct then SMB method produces optimal solution. The correctness of LP formulation depends on the correctness of the design of constraints. There are three constraints involved herein.

A. Capacity constraint on the total flow from all the sources
B. Capacity constraint on the flow of individual source
C. Capacity constraint on the flow of individual commodity

According to Ford Fulkerson theorem [9] capacity constraint of any flow network can be computed by determining the size of its minimum cut. On the other hand, size of the minimum cut can easily be determined by applying any well-known flow maximization method on the network. Therefore above 3 constraints can easily be determined by applying mass balancing method on the respective networks. Now if the design of the respective networks is correct then estimation of constraints and corresponding LP formulation would also be correct. Thus final solution obtained through SMB method is the optimal solution. The following are the proofs that the respective networks designed in the SMB method are correct.

4.1. Lemma 1: Minimum cut of USN (Figure 2) = constraint A

Proof: If all the source nodes are turned into junction nodes and a single source of unlimited capacity joins all those junction nodes through connecting edges of unlimited capacity, i.e., capacity of $E_i = \infty$, where E_i = Set of all the connecting edges then maximum flow through this network represents maximal flow through original multiple source network. This was proved in minimum cut theorem [9]. However simpler explanation is produced here and that explanation is also used later in Lemma-2 and Lemma-3 in support of their proof. Consider a single path network in Figure 4. Now suppose that G_i represents capacity of edge E_i such that

$$G_i = min(G_0, G_1, G_2, \dots, G_n) \qquad (11)$$

According to equation 11 edge E_i has minimum capacity in all the edges from the source S to sink T. Therefore E_i represents minimum cut C_{min} of this network and the maximum flow that could pass through this network cannot be greater than the capacity of edge E_i, i.e.

$$C_{min} = G_i \qquad (12)$$

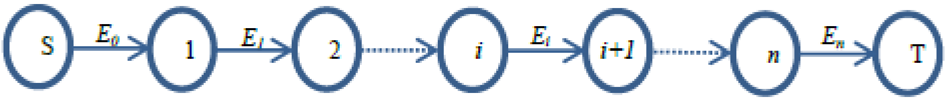


Figure 4. A single path network

Now suppose if the source node in Figure 4 is turned into a junction node and then this node is connected with another source of unlimited capacity through edge E_j as shown in Figure 5.

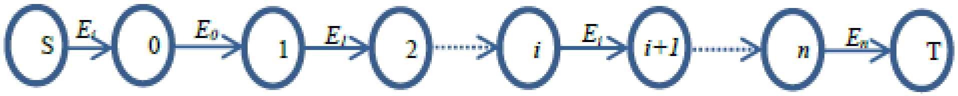


Figure 5. An extended single path network

Now it is very obvious from new network that

$$\begin{cases} C_{min} = G_j & if \quad G_j < G_i \\ C_{min} = G_i & otherwise \end{cases} \qquad (13)$$

Now if edge E_j is considered of unlimited capacity i.e., $G_j = \infty$ then according to condition (13), equation 12 remains true even after the modification in the network.

Since in USN the same modification is incorporated therefore minimum cuts of all paths from sources to sink remain unaffected, and hence minimum cut of overall network remains the same, thus it represents a bottleneck capacity for flow from all the sources in a multi-commodity network.

4.2. Lemma 2: Minimum cut of ISN (Figure 2) = constraint B

Proof: Suppose USN (Figure 2) consists of n connecting edges of unlimited capacity i.e. capacity of $E_i = \infty$ and $E_i \in \{E_1, E_2, E_3, \dots, E_n\}$ then as proved in Lemma-1 any flow maximization algorithm on this network determines maximum capacity constraint on flow from all the sources S_1, S_2, S_3, ….., S_n or determines minimum cut of the network connecting all these sources with the separator/sink.

Now according to definition of ISN as explained in section 2.1, for the ISN_i of source S_i capacity of $E_i = \infty$ and capacity of $E_j = 0$ where $j \in \{1, \ldots\ldots, n/j \neq i\}$, then according to condition 13 minimum cut of all ISN_j emanating from set of sources S_j becomes zero. However minimum cut of ISN_i emanating from source S_i remains unaffected. Hence the ISN_i can be used to determine maximum flow from S_i to sink/separator.

4.3. Lemma 3: Minimum cut of ICN = constraint C

Proof: Here set of edges of unlimited capacity i.e. capacity of $e_i = \infty$ is created, connecting the separator/source with each commodity network. Then by definition in section 2.2, ICN_i for commodity i can be created such that capacity of $e_i = \infty$ and capacity of $e_j = 0$ where $j \in \{1, \ldots\ldots, n/j \neq i\}$.

Now according to condition 13, minimum cut of all ICN_j emanating for set of commodities j becomes zero. However minimum cut of ICN_i emanating for commodity i remains unaffected. Hence the ICN_i can be used to determine maximum flow from source/separator to sink T_i.

5. SOLVED EXAMPLE

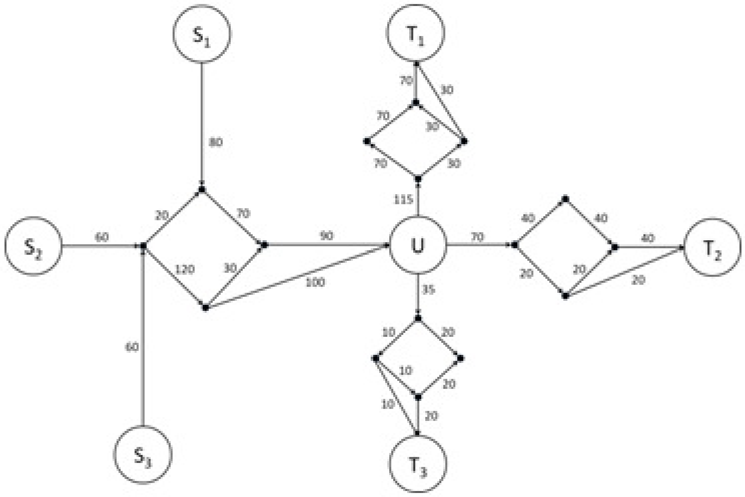


Figure 6. The example of multi-commodity network with a separator

Table 1. Source Configuration

	S_1	S_2	S_3
C_1	0.6	0.5	0.4
C_2	0.3	0.2	0.5
C_3	0.1	0.3	0.1

Figure 6 consists of network with separator. It has three sources S_1, S_2, & S_3. Each source has 3 commodities. Ratio of each commodity in each source is given in Table 1. Consider commodity C_1 as a commodity of interest which needs to be maximized.

The network in Figure 7 represents USN if $E_1 = E_2 = E_3 = \infty$. Since there are 3 sources hence there must be 3 ISNs in Figure 7. By definition of ISN, in section 2.1., in Figure 7, ISN_1

constitutes $E_1 = \infty$ and $E_2 = E_3 = 0$, ISN_2 constitutes $E_2 = \infty$ and $E_1 = E_3 = 0$ and ISN_3 constitutes $E_3 = \infty$ and $E_1 = E_2 = 0$.

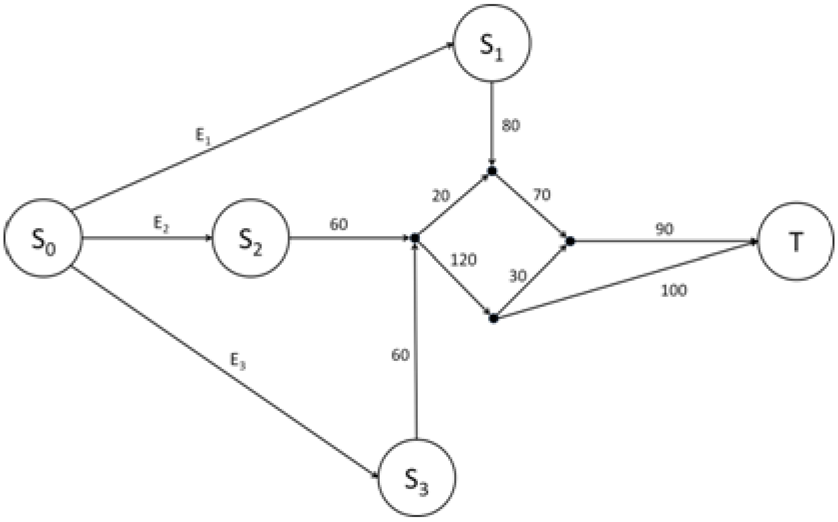


Figure 7. USN and ISNs of the example of Figure 4

Since there are 3 commodities in the network there must be 3 ICNs as shown in Figure 8. By definition of ICN in section 2.2, in Figure 8 ICN_1 constitutes $e_1 = \infty$ and $e_2 = e_3 = 0$, ICN_2 constitutes $e_2 = \infty$ and $e_1 = e_3 = 0$ and ICN_3 constitutes $e_3 = \infty$ and $e_1 = e_2 = 0$.

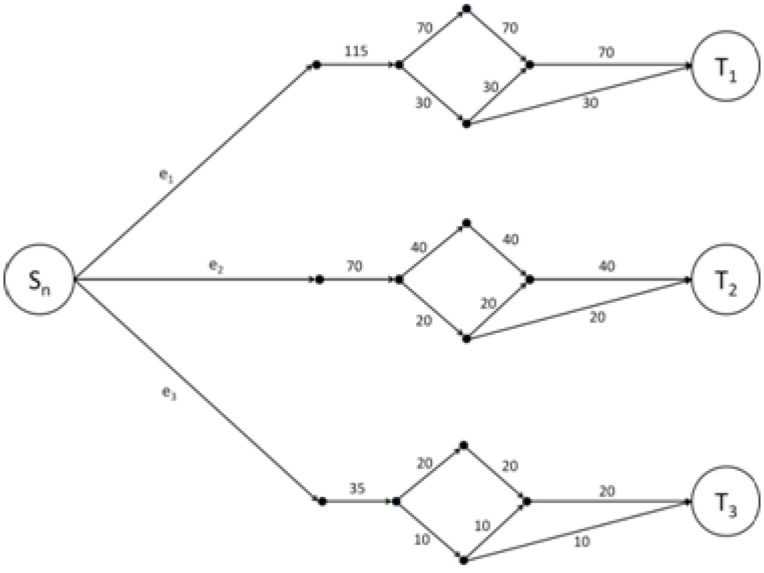


Figure 8. ICNs of the example of Figure 4

With formulation of Figure 7 and Figure 8 step 1 of algorithm has been completed i.e. USN, ISNs and ICNs have been created. In step 2 of algorithm flow is to be maximized in all these networks to determine their minimum cuts to develop LP formulation. Let C_0 be the minimum cut of the USN. After applying flow maximization algorithm on the USN of Figure 7 we get C_0=190. Let

C_{si}, be the minimum cut of ISN$_i$. After applying flow maximization algorithm on ISNs of Figure 7 we get C_{s1}=70, C_{s2}=60 and C_{s3}=60. Let C_{ci} be the minimum cut of ICN$_i$. After applying flow maximization algorithm on ICNs of Figure 8 we get C_{c1}=100, C_{c2}=60 and C_{c3}=30. In step 3 of the algorithm LP formulation is developed from this data as follows.

Maximize the linear function

$$Z = 0.6x1 + 0.5x2 + 0.4x3 \tag{14}$$

Under the constraints

$$\begin{Bmatrix} 0 \ \leq x_1 \leq 70 \\ 0 \ \leq x_2 \leq 60 \\ 0 \ \leq x_3 \leq 60 \end{Bmatrix} \tag{15}$$

$$\begin{Bmatrix} x_1 + \ x_2 + x_3 \leq 190 \\ 0.6x_1 + 0.5x_2 + 0.4x_3 \leq 100 \\ 0.3x_1 + 0.2x_2 + 0.5x_3 \leq 60 \\ 0.1x_1 + 0.3x_2 + 0.1x_3 \leq 30 \end{Bmatrix} \tag{16}$$

Where x_i = flow from source *i*

In Step 4 of the algorithm above linear function is maximized through simplex method and following solution is obtained

$$\begin{Bmatrix} x_1 = 70 \\ x_2 = \ 58.4615 \\ x_3 = \ 54.6154 \end{Bmatrix} \tag{17}$$

In step 5 of the algorithm flow is maximized through the USN by equating *E$_1$*, *E$_2$* and *E$_3$* of Figure 7 with *x$_1$*, *x$_2$* and *x$_3$* respectively.

In step 6 of algorithm, output of each commodity is computed from the results of equation 17 as follows.

$$\begin{Bmatrix} y_1 = 0.6 \times 70 \ + 0.5 \ \times 58.4615 + \ 0.4 \ \times 54.6154 = 93.07691 \\ y_2 = 0.3 \times 70 \ + 0.2 \ \times 58.4615 + \ 0.5 \ \times 54.6154 = 60 \\ y_3 = 0.1 \times 70 \ + 0.3 \ \times 58.4615 + \ 0.1 \ \times 54.6154 = 30 \end{Bmatrix} \tag{18}$$

Where y_i = flow of commodity *i*

In step 7 of the algorithm we maximize flow through the each ICN by equating *e$_1$*, *e$_2$* and *e$_3$* of Figure 8 with *y$_1$*, *y$_2$* and *y$_3$* of equation 18 respectively.

In the final step the USN obtained in step 5 is joined with ICNs obtained in step 7 and all the added edges *E$_1$*, *E$_2$* and *E$_3$* along with the universal source node is removed to represent actual network with maximized flow of COI. The final solution is shown in Figure 9.

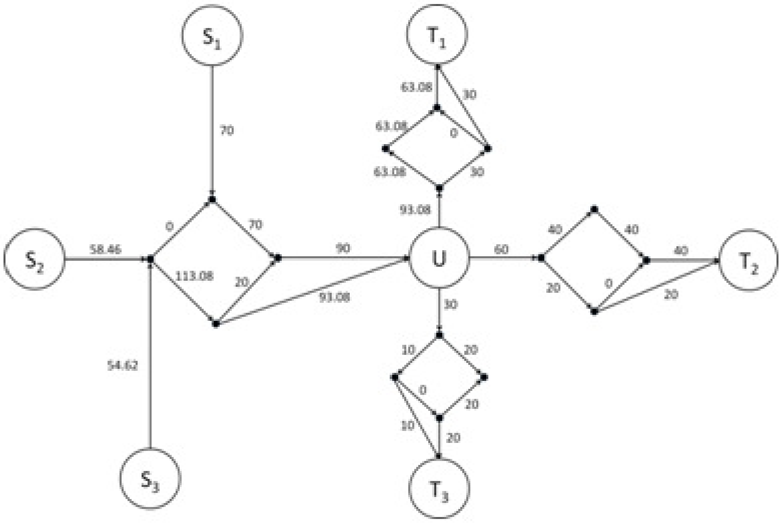


Figure 9. Final solution of Network with Separator

6. Complexity Analysis

The complexity ω of mass balancing theorem (MBT) has already been established [31] and i.e., of order $O(m^2 - m)$ where m is the number of edges. It can be seen that MBT procedure is applied twice on USN (Step 2, 5) twice on ICN (Step 2, 7), and once on ISN (Step 2). Therefore total complexity of MBT procedure on network with separator is given by:

$$\omega = (2 + s)(m_1^2 - m_1) + 2t(m_2^2 - m_2) \tag{19}$$

where

m_1 = number of edges in USN and ISN
m_2 = average number of edges in ICN
t = the number of commodities
s = number of sources

In case of very large networks having tens of thousands of edges, the number of sources and number of commodities would become meaningless thus complexity reduces to:

$$\omega = (m_1^2 - m_1) + (m_2^2 - m_2) \tag{20}$$

Since this complexity can never be greater than $O(m^2 - m)$, where m is the total number of edges in the multi-commodity network with a separator i.e. $m = m_1 + m_2$. Therefore, complexity of MBT method for network with separator stands the same as that of standard network with one source and one sink. However to compute overall complexity of SMB method, complexity of simplex method should also be added into this. However simplex method depends only on number of commodities rather than the network size, hence again it will become meaningless in large networks. Therefore complexity of SMB method stands at $O(m^2 - m)$ only.

7. Conclusion and Future Work

This paper introduces a problem of flow maximization through a multi-commodity network with a separator and describes its applications in oil and gas development fields and mining industry among others. The paper also presents a method to maximize flow through this network, hereby referred to as the Simplex Mass Balancing (SMB) method. The proposed method uses combination of two most important but distinct branches of computer science i.e. linear programming and graph theory. The computational cost of the SMB method is small because linear programming is applied only on the flow input from the sources and flow output from the separator not including the network, while mass balancing theorem is applied on the networks only for flow maximization to determine the constraints needed for LP formulation. This combination has resulted in optimal solution with very less computational cost. The proof of optimality has also been formulated. The future direction of proposed work can be its extension to more complex problems like network with multiple separators.

Acknowledgements

The authors are grateful for financial support from Innovate UK and Route Monkey Ltd via KTP Partnership number 9839.

References

[1] R. K. Ahuja, T. L. Magnanti and J. B. Orlin, (1988) Network flows. Working paper: OR 185-88. Sloan School of Management, MIT, Cambridge, MA.

[2] D. Sheela and C. Chellamuthu, (2012) Protection in Optical Mesh Networks - A Cost Effective Strategy Based on Topological Constraints, Malaysian Journal of Computer Science (ISSN 0127-9084) 25(1), 38-55.

[3] H. Cho, M. Lee and G. Hwang, (2014) A cross-layer relay selection scheme of a wireless network with multiple relays under Rayleigh fading, Journal of Industrial and Management Optimization (JIMO) 10(1), 1-19.

[4] Y. K. Lin and C. F. Huang, (2013) Stochastic Flow Network Reliability with Tolerable Error Rate, Quality Technology and Quantitative Management. 10(1), 57-73.

[5] Z. Ursani, (2014) Computing availability for redundant flow systems. Optimization Letters, 8(2), 715-725.

[6] A. Bhan and E. Mjolsness, (2006) Static and Dynamic Models of Biological Networks. Complexity. Willey Inter Science © 2006 Wiley Periodicals, Inc., 11(6).

[7] M. Masin, M. O. Pasaogullari and S. Joshi, (2007) Dynamic scheduling of production-assembly networks in a distributed environment. IIE Transactions 39, 395–409.

[8] M. Lytras, L. Zhuhadar, J. X. Zhang and E. Kurilovas, (2014) Advances of Scientific Research on Technology Enhanced Learning in Social Networks and Mobile Contexts: Towards High Effective Educational Platforms for Next Generation Education, J.UCS 20(10), 1402–1406.

[9] L. R. Ford and D. R. Jr. Fulkerson, (1956) Maximal flow through a network. Canadian Journal of Mathematics, 8, 399-404.

[10] P. Elias, A. Feinstein, C. E. Shannon, (1956) A Note on the Maximum Flow Through a Network. IRE Transactions on Information Theory, 117-119.

[11] S. L. Gary (1958) Single point of failure: the ten essential laws of supply chain risk management. John Wiley & Sons, Inc. ISBN 978-0-470-42496-4

[12] H. Okamura and P. D. Seymour, (1981) Multicommodity flows in planar graphs. J. Combin. Theory, Ser. B 31, 75–81.

[13] T. Leighton and R. Satish, (1999) Multicommodity Max-Flow Min-Cut Theorems and Their Use in Designing Approximation Algorithms. Journal of the ACM, Vol. 46, 787-832.

[14] E. A. Dinic, (1970) Algorithm for solution of a problem of maximum flow in networks with power estimation. Soviet Math. Dokl. 11, 1277-1280.

[15] J. Edmonds and R. M. Karp, (1972) Theoretical improvements in algorithmic efficiency for network flow problems. Journal of the ACM, 19, 248-264.

[16] R. K. Ahuja and J. B. Orlin, (1989) A fast and simple algorithm for the maximum flow problem. Operations Research, 37(5), 748-759.

[17] R. K. Ahuja, and J. B. Orlin, (1991) Distance-directed augmenting path algorithms for maximum flow and parametric maximum flow problems. Naval Research Logistics, 38, 413-430.

[18] H. N. Gabow, (1985) Scaling algorithms for network problems. Journal of Computer and System Sciences, 31, 148-168.

[19] J. B. Orlin, R. K. Ahuja, (1987) New distance-directed algorithms for maximum flow and parametric maximum flow problems. Working Paper 1908-87, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

[20] A. W. Boldyreff, (1955) Determination of the maximal Steady State Flow of Traffic through a Railroad Network. JORSA, 3(4), 443-465.

[21] A. V. Karzanov, (1974) Determining the maximal flow in a network by the method of pre-flows. Soviet Mathematics Doklady, 15, 434-437.

[22] R. V. Cherkasky, (1977) Algorithm for construction of maximum flow in networks with complexity of O(V2√E) operation. Mathematical Methods of Solution of Economical Problems, 7, 112-125 (in Russian).

[23] V. M. Malhotra, M. P. Kumar and S. N. Maheshwari, (1978) An O(V3) Algorithm for Finding Maximum Flows in Networks. Information Processing Letters, 7, 277-278.

[24] Z. Galil, (1980) O(V5/3E2/3) algorithm for the maximum flow problem. ActaInformatica, 14, 221-242.

[25] R. E. Tarjan, (1986) Algorithms for Maximum Network Flow. Mathematical Programming, 26, 1-11.

[26] A. V. Goldberg, (1985) A new max-flow algorithm. Technical Report MIT/LCS/TM-291, Laboratory for Computer Science, MIT, Cambridge, Mass.

[27] A. V. Goldberg, R. E. Tarjan, (1988) A New Approach to the Maximum-Flow Problem. Journal of the Association for Computing Machinery, 35(4), 921-940.

[28] R. E. Tarjan, (1984) A simple version of Karzanov's blocking flow algorithm. Operations Research Letters, 2, 265-268.

[29] D. S. Hochbaum, (2008) The Pseudo-flow Algorithm. A new algorithm for the maximum flow problem. Operations Research (Informs) 56(4), 992-1009.

[30] J. Dong, L. Wei, C. Cai and Z. Chen, (2009) Draining algorithm for the maximum flow problem. International Conference on Communications and Mobile Computing.

[31] Z. Ursani, (2012) Introducing Mass Balancing Theorem for Network Flow Maximization. International Journal of Industrial Engineering Computations 3, 843-858.

[32] M. Kumral, (2004) Genetic algorithms for optimization of a mine system under uncertainty, Production Planning & Control: The Management of Operations, 15(1), 34-41.

[33] Z. Ursani, (2014) Iterative Mass Balance Method for Multi-Commodity Maximization Problem. Production Planning & Control: The Management of Operations. 25(7), 592-602.

[34] G. B. Dantzig, (1951) Maximization of a Linear Function of Variables Subject to Linear Inequalities. Chapter XXI in "Activity Analysis of Production and Allocation", edited by T. C. Koopmans, Cowles Commission Monograph 13, John Wiley & Sons, New York.

[35] F. Harary, (1969) Graph Theory, Reading, MA: Addison-Wesley.

AUTHORS

Dr. Ziauddin Ursani graduated in Civil Engineering from Mehran University of Engineering and Technology Jamshoro, Sindh, Pakistan in 1989. He obtained Postgraduate Diploma in Environmental Engineering from the same university in 1999. Dr. Ursani did his masters in Information Technology from Hamdard University Karachi, Sindh, Pakistan in 2004 under the award of Science and Technology Scholarship. He received his PhD(Computer Science) in 2009 from Australian Defence Force Academy, University of New South Wales, Australia under the award of prestigious University College Postgraduate Research Scholarship (UCPRS).

He has worked in several postdoctoral projects in UK universities including Oxford Brookes University (Leverhulme Trust Project), University of Salford (Engineering and Physical Sciences Research Council 'EPSRC' Project) and University of Warwick (Department of Energy and Climate Change 'DECC' Project). Presently he is working as a Knowledge Transfer Partnership (KTP) Associate in the Innovate UK Technology Strategy Board Project awarded to Heriot Watt University and Route Monkey Limited. He is engaged in the development of route planning software for the company. Optimisation is his main research interest.

Ahsan Ahmad Ursani was born in Pakistan in 1972. He received the B.Eng. degree in electronics from Mehran University of Engineering and Technology (MUET), Jamshoro, Sindh, Pakistan, in 1995, and the Ph.D. degree in signal and image processing from the National Institute for Applied Sciences (INSA), Rennes, France, in 2008. He has been working since 1995 as a faculty member at MUET and the Chairman of the Department of Bio-Medical Engineering at MUET. His research interests include Remote Sensing and Speech Processing, Biomedical Engineering, Medical Imaging, Digital Signal and Image Processing, and Mathematical Optimization. He worked as a visiting research associate at the International Centre for Theoretical Physics (ITCP), Trieste, Italy, in 1998 and 2003.

Currently he is working as Associate Professor in the faculty of Mathematics and Computer Science at the South Asian University New Delhi, India, where he teaches Discrete Mathematics, Research Methodology, Image Processing, and Embedded Systems Design.

Professor David Corne leads the Intelligent Systems Lab, and is Directtor of Enterpirse, Impact and Innovation, at Heriot-Watt University's Department of Computer Science. He also co-leads the SICSA Data Science research theme (SICSA is the Scottish pool of computer science depts, comprising ~1,000 researchers). He has 25 years' experience in optimization, data science, and machine learning, with extensive industry experience, over 200 publications, three patents and two spinouts. He increasingly focusses on applications in the environmental, energy and well-being sectors. Among several techniques now commonly used in industry for solving large-scale problems in optimization and data analytics, he has developed and co-developed new ideas in logistics, multi-criterion and large-scale optimization. His current focus, working closely with Route Monkey Ltd, is on intelligent algorithms for energy and mobility optimization.

INTENTIONAL BLANK

A STUDY AND IMPLEMENTATION OF THE TRANSIT ROUTE NETWORK DESIGN PROBLEM FOR A REALISTIC URBAN CASE

M. Kalochristianakis[1] and D. Kosmopoulos[2]

[1]Department of Informatics Engineering,
Technological Educational Institution of Crete, Heraklion, Greece
kalohr@staff.teicrete.gr
[2]Department of Cultural Heritage Management and New Technologies,
University of Patras, Agrinio
dkosmo@upatras.gr

ABSTRACT

The design of public transportation networks presupposes solving optimization problems, involving various parameters such as the proper mathematical description of networks, the algorithmic approach to apply, and also the consideration of real-world, practical characteristics such as the types of vehicles in the network, the frequencies of routes, demand, possible limitations of route capacities, travel decisions made by passengers, the environmental footprint of the system, the available bus technologies, besides others. The current paper presents the progress of the work that aims to study the design of a municipal public transportation system that employs middleware technologies and geographic information services in order to produce practical, realistic results. The system employs novel optimization approaches such as the particle swarm algorithms and also considers various environmental parameters such as the use of electric vehicles and the emissions of conventional ones.

KEYWORDS

Public transport network, environmental optimization, particle swarm optimization, geographic informational systems, middleware

1. INTRODUCTION

The problem of optimizing the use of resources with respect to the environmental impact has been an area of focus during the last decade [1] [2]. The design of a public transportation network is a complex optimization problem, which involves a variety of design parameters (route structure, frequencies, vehicle types, etc) and assumptions on demand patterns, travel behavior and so on. Indeed, the associated Transit Route Network Design Problem (TRNDP) has been a topic of interest for over 40 years. The combinatorial nature of the TRNDP and the difficulty to formulate it analytically have resulted numerical optimization as the primary means of approaching the solutions over the last years. A review of the recent literature exhibits a variety

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 59–65, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51506

of relevant techniques that consider routes, frequencies and other network parameters, based on preset objective functions, which are to be optimized. Widely used approaches include Genetic Algorithms (GA) [3], Simulated Annealing [4] and Ant Colony Optimization [5] besides others.

The Particle Swarm Optimization (PSO) is one of the most effective evolutionary algorithms inspired from social behavior of animals [6]. Its simplicity and efficiency makes this algorithm very popular. Due to these advantages, the PSO algorithm has been applied to many domains such as medical diagnosis, grid scheduling, robot path planning and computer vision. This algorithm is capable of solving problems with continuous search spaces, while some problems have discrete search spaces. The binary version of PSO (BPSO) was proposed by [7]. The TRNDP belongs to the discrete problems and probably this is the reason that the PSO algorithm has not been applied to this problem so far.

The rest of the paper presents the formulation of the TRNDP problem so that PSO optimization procedures can be used to approach its solution. The paper is structured as follows: section 2 analyses the formulation of the problem. Section 3 describes the component architecture of the proposed framework. Section 4 presents the conclusions of this work and the future perspective.

2. Problem Formulation

This work has been based on the assumptions that there is a fixed number of *S* bus stops, a fixed number of bus lines *L* and that the bus lines have a maximum number of *s* bus stops. The solution is represented by a binary two dimensional matrix of *L* rows and s columns. The *l*-th row represents the *l*-th bus line. A “1” in position (l, *σ*) represents that the *l*-th bus line goes through the *σ*-th bus stop, while a “0” represents that the bus line does not include the bus 2stop. The solution must be in vector form, therefore, we vectorize the 2D matrix to formulate the hereafter mentioned as “*LN*” vector with *L* × *s* elements. To the previous vector we also have to append bits to encode bus frequencies per line (number of bits depends on what is the maximum bus frequency) hereafter denoted as “*f*” and whether the line is operated by electric or conventional bus hereafter denoted as “*G*”. The following step is to minimize the objective function given by:

$$\begin{gathered} minZ = w_1 D_u(\bar{L}\bar{N},\bar{f}) + w_2 T(\bar{L}\bar{N},\bar{f}) \\ + w_3 e(\bar{L}\bar{N},\bar{f},\bar{G}) + w_3 N_{cs}(\bar{G}) + w_5 V_c(\bar{L}\bar{N},\bar{f},\bar{G}) + \\ w_6 D_e(\bar{L}\bar{N},\bar{f},\bar{G}) \end{gathered} \qquad (1)$$

where D_u is the unsatisfied passenger demand (not served under maximum transfers), *T* the average travel time, *e* the pollution emissions N_{cs} the number of charging stations, V_c the required number of conventional vehicles and V_e the required number of electric vehicles. The weights w_1 - w_6 are defined according to the policy we want to implement or according to values that can be statistically estimated. All the above quantities are straightforward to compute given *LN*, *F*, *G*. However, the question is: given the solution vector how do we define the sequence of the bus stops? Clearly the solution vector does not really capture the sequence in which the bus stops are visited. This is done deliberately and is one of our major contributions, because we significantly reduce the solution space. E.g. for *S* = 50 and *s* = 10 the number of possible permutations in which we are seeking optimum is ~10^{16} , while if we ignore the permutations as in our method this reduces the search space to ~10^{11} .

The answer is given by assuming that the next bus stop is the one closest to the current one. In other words we need to define the path that covers all the bus stops and at the same time has the minimum possible length. The answer to this problem is given by the Hamiltonian path, which solves exactly this problem [8].

3. THE BINARY PSO ALGORITHM

The binary version of PSO (BPSO) was proposed by [7]. The continuous and binary versions of PSO are distinguished by two different components: the transfer function and the different position updating procedure. The transfer function is used to map a continuous search space to a binary one, and the updating process is designed to switch positions of particles between 0 and 1 in binary search spaces. Several solutions have been proposed to the problem of getting trapped in local minima, e.g., [10], [11]. In [12], two different families of transfer functions, v- shaped and s-shaped were investigated. Let's start from the continuous PSO. Each particle *i* at time *t* corresponds to a single solution *$x_i(t)$*. To evolve towards a better solution the particle has to consider the current position, the current velocity *$v_i(t)$*, the distance to their personal best solution, *pbest*, and the distance to the global best solution, *gbest*. This is formulated as follows:

$$v_i(t+1) = w * v_i(t) + c_1 * r_1 * (pbest - x_i(t)) + c_2 * r_2 * (gbest - x_i(t)) \qquad (2)$$

where w is a weighting function, r_1, $r_2 \in [0,1]$ are random numbers and c_1, c_2 are acceleration coefficients. In the next iteration the particle will evolve to:

$$= x_i(t) + v_i(t+1) \qquad (3)$$

In binary space, due to dealing with only two numbers (''0'' and ''1''), the position updating process cannot be performed using eq. (3). Therefore, another definition of velocities is needed for changing positions from ''0'' to ''1'' or vice versa. This can be done by redefining the velocity to be the probability of a bit taking the value *0* or *1*. A sigmoid transfer function as in eq. (4) was employed in [1] to transform all real values of velocities to probability values in the interval [0,1].

$$T(\ v_i^k(t)\ ) = \frac{1}{1 + e^{-\ v_i^k(t)}} \qquad (4)$$

where $v_i{}^k(t\ )$ indicates the k-th dimension of the velocity vector. Then the position vectors are updated according to the following:

$$x_i^k(t+1) = \begin{vmatrix} 0, & if\ r < T(v_i^k(t+1)) \\ 1, & if\ r \geq T(v_i^k(t+1)) \end{vmatrix} \qquad (5)$$

where r is a random number in the interval [0, 1]. Variations of this strategy have been proposed in [10], [11], [12].

4. Component Architecture and Algorithmic Approach

The design of our case study is based on the combination of presentation technologies, middleware and computational analysis, namely: HTML, Javascript, Google Maps API at the presentation tier, Hypertext Preprocessor (PHP), the known dynamic programming language for the middleware and Octave / Matlab for the computational analysis back-end. More specifically, the systems includes a graphical web interface capable of displaying the graph of the problem realistically and in real time through the programming interface of Google Maps engine management. The graph presented by Google Maps is processed by means of a computational analysis module which implement the TRNDP solver based on the Octave and Matlab environments. Middleware logic, is capable to handle requests directed towards the graphical interface, direct them to the computational analysis module and return results in appropriate form to be presented by the maps presentation engine.

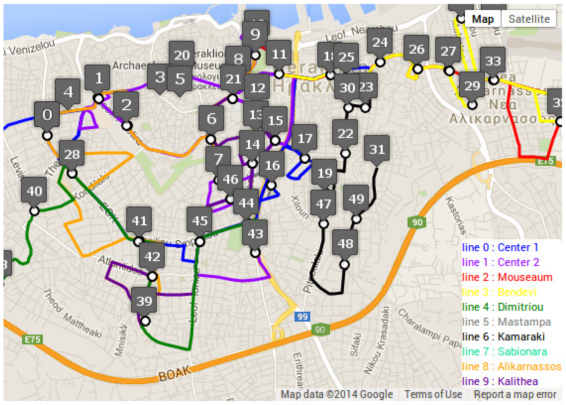
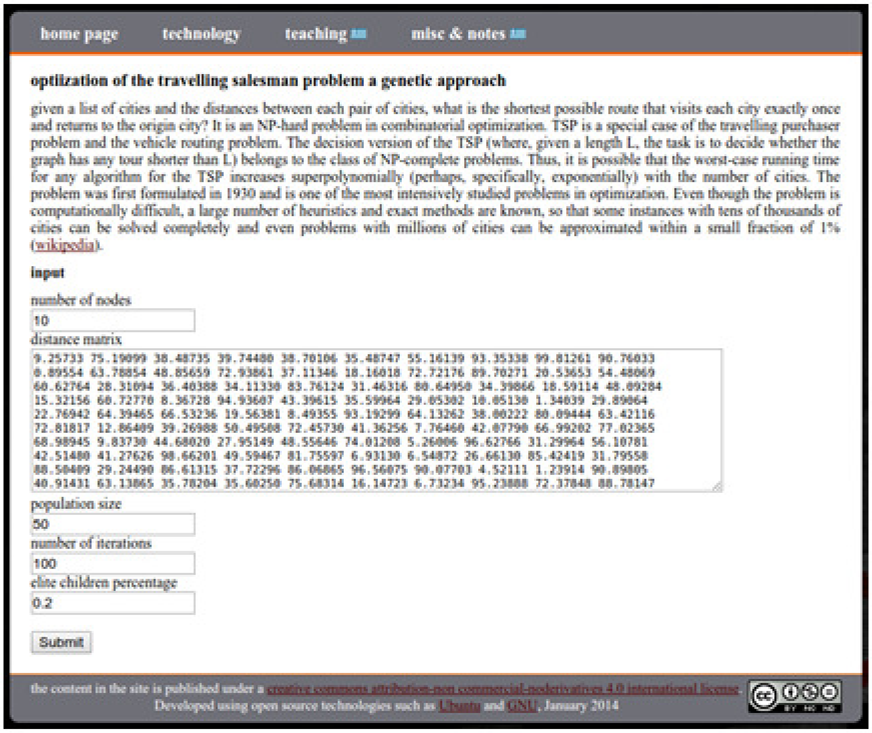


Fig. 1: an initial configuration of the system that is, a realistic selection of graph nodes and routes can be displayed on Google Maps and also information about distances and traffic can be retrieved and passed to the computational analysis system.

The Google Maps map management and presentation engine offers programming interfaces that are compatible with various programming languages. For the purposes of our work, we took advantage of the Javascript programming [9]. Necessary conditions for the use of the interface has been the knowledge of web technologies and principles of object-oriented programming design mode. This platform was chosen because it dominates the market of GIS and provides free, stable and reliable access. Google maps also offer interesting features such as real time traffic support, carbon dioxide emissions estimations, beside others. The use of the programming interfaces of the platform service is offered by means of subscription and the acquisition of appropriate application programming interface (API) keys that allow the service to monitor usage. Typical facilities include the DistanceMatrixService offering distance calculation service between start and destination nodes, DirectionsService offering directional calculation service between one or more locations, DirectionsRoute service offering route calculation between departure and destination which contains the sections of the route, among others, Map objects capable to illustrate maps. Features of the service include vehicle type specification, travel modes that is bicycling, driving, transit, or walking. As a commercial product, the Google Maps API allows limited use when not in subscription mode. In this context the design of our system took into account the respective restrictions. When the design took place the former allowed 25,000 map loads per day for 90 consecutive days, recovering 100 elements each performed search (query), recovering 100 elements per 10 seconds, recovery of 2,500 items per 24 hours. It is worth noting that requests are also subject to rate limits. The design of the system took into account all the above restrictions in order for the system to be capable to represent nodes (stops) as points in Google Maps, represent of realistic routes ie routes that take account of actual characteristics as e.g. one-way. An initial configuration of the system is illustrated in fig. 1.

GNU Octave is a high-level programming language, primarily intended for numerical computations. It offers a command line interface for solving linear and nonlinear problems numerically, and for performing other numerical experiments using an interpreted language mostly compatible with the well known MATLAB platform by Mathworks. It can also be used as a language oriented script execution. Octave is free software, distributed under the terms of the GNU General Public License. Besides its use for desktop computers that is, for personal scientific calculations, Octave is also used in academia and industry. Its features include that it is written in C ++ and uses the standard C ++ library, it uses an interpreter to execute the script language and it is expandable with the use of dynamic parts (modules). Versions 3.8.0 and later include a graphical user interface (GUI), other than the traditional command line interface (CLI). The architecture of our solution employs open source PSO packages compatible with Octave.



optilization of the travelling salesman problem a genetic approach

given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city? It is an NP-hard problem in combinatorial optimization. TSP is a special case of the travelling purchaser problem and the vehicle routing problem. The decision version of the TSP (where, given a length L, the task is to decide whether the graph has any tour shorter than L) belongs to the class of NP-complete problems. Thus, it is possible that the worst-case running time for any algorithm for the TSP increases superpolynomially (perhaps, specifically, exponentially) with the number of cities. The problem was first formulated in 1930 and is one of the most intensively studied problems in optimization. Even though the problem is computationally difficult, a large number of heuristics and exact methods are known, so that some instances with tens of thousands of cities can be solved completely and even problems with millions of cities can be approximated within a small fraction of 1% (wikipedia).

input

number of nodes: 10

distance matrix

```
9.25733 75.19099 38.48735 39.74480 38.70106 35.48747 55.16139 93.35338 99.81261 90.76033
0.89554 63.78854 48.85659 72.93861 37.11346 18.16018 72.72176 89.70271 20.53653 54.48069
60.62764 28.31094 36.40388 34.11330 83.76124 31.46316 80.64950 34.39866 18.59114 48.09284
15.32156 60.72770 8.36728 94.93607 43.39615 35.59964 29.05302 10.05130 1.34039 29.89064
22.76942 64.39465 66.53236 19.56381 8.49355 93.19299 64.13262 38.00222 80.09444 63.42116
72.81817 12.86409 39.26988 50.49508 72.45730 41.36256 7.76460 42.07790 66.99202 77.02365
68.98945 9.83730 44.68020 27.95149 48.55646 74.01208 5.26006 96.62766 31.29964 56.10781
42.51480 41.27626 98.66201 49.59467 81.75597 6.93130 6.54872 26.66130 85.42419 31.79558
88.50409 29.24490 86.61315 37.72296 86.06865 96.56075 90.07703 4.52111 1.23914 90.89805
40.91431 63.13865 35.78204 35.60250 75.68314 16.14723 6.73234 95.23888 72.37848 88.78147
```

population size: 50

number of iterations: 100

elite children percentage: 0.2

the content in the site is published under a creative commons attribution-non commercial-noderivatives 4.0 international license.
Developed using open source technologies such as Ubuntu and GNU, January 2014

The architectural components and their integration was a challenge for the implementation of the system. Google maps was a valuable and suitable solution since it offers unique functionality, satisfactory front-end interface and sufficient, usable APIs. Matlab is both capable and efficient to execute the algorithmic logic but, by design, not suitable for the middle tier of the platform; Octave on the contrary is very easy to integrate but does not support either PSO or GA to the desirable extent. In order to produce realistic, applicable solutions the system needs to produce meaningful routes; to this end they needed to be optimized with respect to their total distance. Thus, besides optimizing with respect to the objective function analyzed above, it was decided to recover the shortest possible route that visits each node exactly once and returns to the origin. The aforementioned sentence is an expression of the Traveling Salesman problem [13] the well known non-deterministic polynomial time (NP) hard problem. This logic also needed to run in the middle tier.

Fig. 2: in order to ensure that bus routes are optimal with respect to traveling distance the systems solves the traveling salesman problem using a platform independent implementation of a genetic algorithm [14], illustrated in the figure.

5. Conclusions and Future Work

The current paper outlines the engineering and algorithmic design of the DIANNA system that aims to solve the TRNDP problem using a PSO approach. The objective function of the optimization algorithm is very similar to [3]; it ultimately aims to produce solutions that optimize environmental parameters that is, vehicle emissions and vehicle types (electrical or conventional) besides more typical parameters of the problem such as distances, bus frequencies, demand. The design is innovative since the formulation of the solution is binary, designed to facilitate easy manipulation. Also, the architecture of the informational system is designed to interact with well known GIS services, relies on middleware logic that executes the optimization and presents the results using web technologies. Fig. 2 illustrates the implementation of a genetic algorithm that solves the traveling salesman problem; the implementation relies in platform independent server-side logic that can be called by any component of the system.

In the near future we expect to produce experimental results that exhibit realistic solutions for the case of Heraklion, Crete and, since the formulation of the problem allows it, we also expect to investigate the extent to which environmental policies can be applied that is, find optimal or next to optimal values for parameters w_1 to w_6 that correspond to minimum environmental costs.

ACKNOWLEDGEMENTS

This work is part of the DIANA project, funded by the Operational Program "Education and Lifelong Learning", action Archimedes III, co-financed by the EU and National funds (National Strategic Reference Framework 2007-2013).

References

[1] Y. Jang, Y. Ko, "System architecture and mathematical model of public transportation system utilizing wireless charging electric vehicles", Intelligent Transportation Systems, 15th Internationalnal IEEE Conference on, pp.1055-1060, 2012

[2] T.H. Ortmeyer, P. Pillay, "Trends in transportation sector technology energy use and greenhouse gas emissions," Proceedings of the IEEE, vol.89, no.12, pp.1837-1847, 2001

[3] M. Pternea, K. Kepaptsoglou, and M. Karlaftis, "Sustainable urban transit network design", Transportation Research Part A, vol. 77, pp. 276–291, 2015

[4] F. Zhao and X. Zeng, "Optimization of transit route network, vehicle headways and timetables for large-scale transit networks", European Journal of Operational Research, vol. 186, no. 2, pp. 841–855, 2008

[5] J. J. Blum and T. V. Mathew, "Intelligent agent optimization of urban bus transit system design", Journal of Computing in Civil Engineering, vol. 25, no. 5 pp. 357–369, 2010

[6] R. Eberhart, J. Kennedy, A new optimizer using particles swarm theory, in: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995.

[7] J. Kennedy, R. Eberhart, "A discrete binary version of the particle swarm algorithm", in: Proceedings of the IEEE International Conference on Computational Cybernetics and Simulation, 1997.

[8] N. Biggs, "T. P. Kirkman, mathematician", The Bulletin of the London Mathematical Society vol. 13, no. 2, pp. 97–120, 1997

[9] Google MAPS documentation for the Javascript application programming interface https://developers.google.com/maps/documentation/javascript/, accessed on August 2015

[10] S. Lee, S. Soak, S. Oh, W. Pedrycz, M. Jeon, "Modified binary particle swarm optimization", Progress in Natural Science, iss. 18, vol. 9, pp. 1161–1166, 2008

[11] L. Chuang, S. Tsai, C. Yand, "Improved binary particle swarm optimization using catfish effect for feature selection", Expert Systems with Applications, iss. 38, pp. 12699–12707, 2011

[12] S. Mirjalili, A. Lewis, "S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization", Swarm and Evolutionary Computation, pp. 91–14, 2013

[13] M. Tasgetiren, P. Suganthan, P. Quan-ke, L. Yun-Chia, "A genetic algorithm for the generalized traveling salesman problem", IEEE Congress on Evolutionary Computation, pp. 2382 – 2389, 2007

[14] M. Kalochristianakis, http://83.212.103.151/~mkalochristianakis/techNote s/travelling.php, accessed on September 1, 2015

AUTHORS

Michael Kalochristianakis

Michael is currently an associate lecturer and researcher at the Technological Educational Institution of Crete (TEIC), at the Department of Informatics Engineering. His interests include but are not limited to technical project management and project management in general, innovative IT systems and services, software development methodologies, multimedia technologies and algorithms, web engineering, object oriented design, architectures, frameworks and patterns, in-band management and remote management, web and portal development, open source technologies, green technologies and energy conservation. Michael holds a Doctorate degree in Computer Engineering and Informatics, a Masters degree in Computer Science and a Diploma in Electrical Engineering and Computer Technology.

Dimitrios Kosmopoulos

Dimitrios Kosmopoulos is currently Assistant Professor at the University of Patras, Department of Cultural Heritage and New Technologies. Previously he was at Technical Educational Institute of Crete - Department of Informatics Engineering and before that, he was at Rutgers University, Computer Science Department, CBIM Lab and at the University of Texas at Arlington, Department of Computer Science and Engineering. He has been a research scientist in the Computational Intelligence Laboratory of the Institute of Informatics and Telecommunications in the NCSR Demokritos. He was also affiliated with the Department of Electrical and Computer Engineering of the National Technical University of Athens (Greece). He was employed as a researcher and developer for various companies and institutions.

INTENTIONAL BLANK

DEVELOPING MULTITHREADED DATABASE APPLICATION USING JAVA TOOLS AND ORACLE DATABASE MANAGEMENT SYSTEM IN INTRANET ENVIRONMENT

Raied Salman

Computer Information Science,
American College of Commerce and Technology – Falls Church, VA, U.S.A.
raied.salman@acct.edu

ABSTRACT

In many business organizations, database applications are designed and implemented using various DBMS and Programming Languages. These applications are used to maintain databases for the organizations. The organization departments can be located at different locations and can be connected by intranet environment. In such environment maintenance of database records become an assignment of complexity which needs to be resolved. In this paper an intranet application is designed and implemented using Object-Oriented Programming Language Java and Object-Relational Database Management System Oracle in multithreaded Operating System environment.

KEYWORDS

Intranet, Multithreads, OOP, ORDBMS, JDBC, Applets, Oracle, Java Programming Language.

1. INTRODUCTION

The Intranet technology has opened new areas of research for business application designers and implementers. The application is designed using System Development Life Cycle (SDLC) methodology [1,2,4]. The database can be stored on a database server using Oracle Database Management System and can be processed using Java Programming language [5,6,7,8]. Java is an object-oriented programming language. The peoples who have used structured programming languages C, PASCAL etc. has to refuel their programming power to accept object-based programming such as Java and C++ [10] etc. It is very difficult to decide which programming language will lead the application for development in the future. In this paper the basic concepts and tools are discussed which can be used to implement business applications in an intranet environment.

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 67–79, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51507

2. APPLICATION INFRASTRUCTURE FOR INTRANET ENVIRONMENT

It is understood that each database application has to apply four basic functions, *INSER*T, *UPDATE*, *RETRIEVE* and *DELETE* on database records [1,2,3,4]. A database schema is developed using analysis and design techniques. After further refinement this schema is implemented using a specified RDBMS such as ORACLE [12,13,14] on the ORACLE Server. This schema always reflects the data requirements of the organization in which it will be implemented. These basic functions can be implemented using RDBMS selected. The main problem is with the processing of the business applications where many more functions are involved in addition to these four basic functions such as new calculations. The languages provided by the DBMS are not process-oriented so the implementer has to look for a language, which can facilitate the process implementations for business applications. Different database development modelling strategies are discussed in this paper.

2.1 Single Tier Database Design Strategy

The earlier business applications were developed using RDBMS based on an integrated model which consists of *user interface code*, application *code*, and *database libraries*. These applications ran only on a mainframe machine connected to terminals, used to make different queries on the databases. Figure 1 illustrates single-tier application infrastructure.

These business applications were simple but inefficient and did not work over Local Area Networks (LANs). This model did not scale, and the *application code* and the *user interface code* were tightly coupled to the database through database libraries. This approach did not allow multiple instances [1,11,12,13,14] of the application to communicate with each other, so there was often problem of contention between instances of the same business application [12,13]. In order to eliminate some of the contentions occurring, the two-tier database design strategy was suggested [1,2,3,4].

2.2 Two-Tier Database Design Strategy

The server technology gave birth to two-tier RDBMS models. Communication-protocol development and extensive use of LANs and WANs [12,13,14,15] allowed the database developers to create an application front-end that typically accessed data through a connection to the back-end server [14,15]. Figure 2. illustrates a two-tier database design, where the client is connected to the server through a socket [5,6,7,8,9,15] connection. The program design method is very carefully used to accommodate all types of changes taking place in database design strategies [10].

Business applications / client programs through user interface send **SQL** requests to the database server. The server responds with the requested data to the business application / client machine with the specified format, after the verification of these requests. The communication between either of them and the server is managed by the library functions provided by the venders / third party software developers [7,9,10,11]. The limitations to this application design are mentioned below.

2.2.1 Limitations of the Two-Tier Database Model

i. Two-tier models are limited by the vendor-provided library [8, 9]. Switching from one database vendor to another requires a lot of modification to the business application code running on the client machine of the two-tier model.

ii. Version control is another issue. Updating the client-side libraries provided by the vendors causes the database applications to be recompiled and redistributed in the organization [9].

iii. Vendors libraries deal with low-level data manipulation. Many basic libraries deal with fetches and updates on a single row or a column. The stored procedures can be used on the server to enhance these operations increasing the complexity of the application [12,13,14].

iv. All the logic required to use and manipulate the data is implemented in the business application on the client machine, creating large client-side runtimes. This creates a fat client [1,2,3,4].

These limitations can be fully / partially removed from the two-tier model by using a three-tier model.

2.3 Three-Tier Database Design Strategy

In this model the client application communicates with an intermediate server that provides a layer of abstraction from the RDBMS. Figure 3. illustrates this model.

The intermediate layer is designed to handle multiple client requests and manage the connection to one or more database servers. The detail for this design model can be found in [1,2,3,4,14].

3. IMPLEMENTATION OF BUSINESS APPLICATION IN INTRANET ENVIRONMENT

Business applications are mission critical applications. These have to be implemented with great care and sense of responsibility. After the analysis of the user requirements for applications, the implementers have to decide in addition to RDBMS, about the programming languages, which provide the functionality of the application with minimal changes and development time if required to install on different platforms. In the present case, Java programming language is selected to implement such application, because it is platform free language [5,6,7,8]. A segment of the Payroll System is implemented using Java programming language and ORACLE database management System.

The relations / tables, which were used to explain the implementation step are given in Appendix A.

In the *Department* ***table,*** Dept_No ***is a primary key, which has unique values for individual records.***

The second table / relation used to implement **one-to-many** relationship is an ***Employee*** table. In the **Employee** table, ***Emp_No*** is a primary key, and **Emp_Dept_No** is a foreign key to create a one-to- many relationship between them.

The relationship between these two tables is represented in Fig. 4

Since a segment of a **Payroll System** is to be implemented in Java programming language, the multithreading programming technique is used [9] to reduce the development time and other resources.

4. DESCRIPTION OF MULTITHREADING TECHNIQUE IN JAVA IMPLEMENTATION

The concurrency or parallelism that computers can perform is implemented through **Operating Systems** primitives available to highly experienced system programmers [5,6,7]. Using Java programming language these primitives are made available to the application programmers too. Each application can contain threads of execution such that each thread being designated a portion of the application that may execute with other threads concurrently. Multiple threading is a powerful capability of Java language not available in C and C++ [5,6,7,8,9]. Java programming includes multithreading primitives as part of the language in the form of classes such as ***Thread***, ***ThreadGroup***, ***ThreadLocal*** and ***ThreadDeat****h* of *the* ***java.lang*** package [5,6,7,8]. There are many constructor methods related to the ***Thread*** class which play an important role in the ***Thread*** class operations [5,6,7,8,9]. The thread life cycle is given in [5,6].

4.1 Connecting to the ORACLE Database System

It is difficult to join two different technologies such as **Java** based on object–orientation and **ORACLE** based on Relations (tables). Tools which are used to establish the connection between these two different technologies for Multithreaded Intranet Windows applications [5, 6] development are given below.

4.2 Java Database Connectivity (JDBC): Application Programming Interface (API)

Java programming language offers several benefits to the developer creating front-end and middle-ware applications for a database server. The platform-independent nature [5,6,7,8,9] and adaptability of Java [6,7,8,9] allows a wide variety of business applications on the client machines to connect to the database systems installed on the servers [6,7]. Enterprise JavaBeans (EJB) provides a very scalable and robust database access and persistent layer [8, 9].

Servlets and **JSP** (Java Server Pages) [8,9] provide an ideal way for *thin web browser clients* or any variety of other **HTTP**-based clients to access database resources [8, 9]. The **JDBC API** is designed to allow the application developers to create Java code that can be used to access almost any relational database without needing to continually rewrite their application code. Java **servlets**, **JSP** pages, Enterprise JavaBeans (**EJB**) and **Java** classes or any other *Java code* can use **JDBC** to connect to the database server [8,9].

4.3 The JDBC API Characteristics

Recently developed Java Development Kit version 1.4 (JDK 1.4) contains JDBC 3.0 API. It is composed of the java.sql and javax.sql packages.

i. The JDBC interface provides application developer with a single API that is uniform and database independent [9]. Its database independence is due to the availability of a set of Java interfaces that are implemented by a driver [9]. The driver is used to translate the standard JDBC calls into specific calls required by the RDBMS it supports [6,7,8,9].

ii. The business application is developed only once, and then moved to the various drivers, it means that application remains the same and only drivers are changed according to the RDBMS [7,8,9] provided by the vendors.

iii. JDBC also provides a means of allowing developers to retain the specific functionality that their database vendor offers.

iv. JDBC allows the application developers to pass query strings directly to the connected driver. These query string may or may not be ANSI SQL compatible. The query depends on the driver.

v. Every Java application (Client or J2EE) that uses JDBC must have at least one JDBC driver, and each driver is specific to the type of RDBMS under consideration [6,7,8].

vi. JDBC is not derived from Microsoft ODBC [7,8,9]

vii. JavaSoft provides a JDBC-ODBC bridge that translates JDBC calls to ODBC calls [6,7,8,9].

In order to connect business applications / client machines to various RDBMS on the servers, through JDBC are discussed below, for various database design strategies.

4.4 Single-tier JDBC Database Design Strategy

In this configuration, a business application can be connected to different database servers through JDBC interface using different drivers provided by their venders. It is illustrated in Appendix A, Figure 5.

4.5 Multi-tier JDBC Database Design Strategy

In this configuration, a middle tier is used to handle protocols and DBMS libraries implemented for the client sides. Through these protocols, business application can be implemented to access the database servers by different venders in parallel or concurrently. The drivers are dependent on the venders whereas the JDBC is independent of the drivers offered by various venders. The details of this configuration is given in [6,7,8,9] and is illustrated in Appendix A, Fig. 6

5. THE JAVA DATABASE CONNECTIVITY (JDBC) INTERFACE LEVELS

The **JDBC** has two levels of **API** interface: Driver Layer and Application Layer, which are discussed below:

i. **Driver Layer:** It handles all types of communications with a specific driver during implementation to the application layer.

ii. **Application Layer:** This layer is used by the business application developer to make calls to the database via SQL queries and retrieve the results to these queries.

The application developer is not concerned with the details of the implementation of these layers. It is necessary to understand the ***Driver layer***, and how some of the objects that are used in the ***Application layer*** are created by the driver in use [1,3,6,8]. Every driver must implement four main interfaces and one class that create connection between the Driver and Application layers.

5.1 The Driver layer and Driver Interface

Each vendor supplies a driver class called ***DriverManager*** class which controls the Application layer through the driver as an interface. ***Driver Manager*** class also performs: loading and unloading of drivers and making connections using drivers. It also performs some functions on database for login and login times out [6].

a- Driver Interface

It is important to note that every **JDBC** application must have at least one **JDBC** driver. This interface permits the ***DriverManager*** and **JDBC** Application **layer** to exist independently of the database being used. This interface implements **JDBC** driver [6,7,8,9]. Drivers use a string referred to as a *URL* with a purpose to separate the application developer from the driver developer. The syntax for such *URL* for **JDBC** driver is given as

String url = jdbc: <subprotocol>:<subname>

Where <*subprotocol*> is the type of the *driver*, and <*subname*> provides the *network-encoded database name on the server*, as in

String url = "jdbc:oracle:Depts"

In this example, the driver type is **oracle** driver, and the subname is a local database host called **Depts**.

The application developer can also include the location of the database host or instance of the database, the specific port, and user information (user-name, user-password) as in the following example:

String url = "jdbc: oracle: thin: @dbserver:1521: infs" ;

In this statement, the name of the driver is ***oracle*** driver, the name of the database server is ***dbserver*,** the port is **1521** and database instance is ***infs*.**

The following two statements describe the user name and password of the user.

String User = "user_name";
String Password = "user_password";

The driver interface has two important methods from practical point of view [6,7,8,9]:

i- ***public Connection connect (String url, String User, String Password) throws SQLException*.**

In order to return the object of ***Connection*** type, the ***String url*** must match the ***url*** of the **JDBC** driver otherwise no connection will be established. The strings ***User*** and ***Password*** are also matched with those stored on the database server, ***dbserver,*** with instance ***infs*** on the thin client with port **1521**. Since it is public method, the object returned can be used by other classes. If these matches are invalid, it will throw an ***SQLException*** indicating that no connection object is returned.

ii- ***public boolean acceptsURL(String url) throws SQLException.***
This method is simply used to check whether the ***url*** is valid or not. If it is not, it will throw an ***SQLException***. It will not establish the connection.
The ***DriverManager*** class calls the ***Driver connect()*** method to obtain the ***Connection*** object which is the starting point for the ***Application Layer***. The ***Connection*** object is used to create ***Statement*** objects that perform queries.

The DriverManager* Class**: As the name indicates this class is used to manage ***JDBC drivers. Public Methods available in this class are:
i- ***public static synchronized Connection getConnection (String url, String User, String Password) throws SQLException***.
This method is used to obtain ***Connection*** object by sweeping through a vector of stored ***Driver*** classes using ***url*** and other parameter values regarding the user of the database and his password. This method is used to find a ***driver*** which returns a **Connection** object. That ***Driver*** class is used for which the driver is found. This method can be used as an overloaded method with different number of arguments.
ii- ***public static synchronized void registerDriver (java.sql.Driver driver) throws SQLException*.**
This method stores the information of the driver interface implementation into a vector of drivers. It also stores information about security Context [7,8], that identifies where the driver came from.
iii- ***Public static void setLogWriter(java.io.PrintWriter out).***
Sets a private static **java.io.PrintWriter** reference to the **PrintWriter** object passed to the method.

b- Registration of Drivers

When ***DriverManager*** class is loaded, a static code of this class is executed to load ***jdbc.drivers***.
jdbc.drivers property can be used to define a list of colon-separated driver class names such as:
***jdbc.drivers = oracle.jdbc.driver.OracleDriver*;**
Each driver name is also a class name [6,7,8], this means that class name and driver name are the same, for example, ***oracle.jdbc.driver*** is both a driver name and a class name. The ***DriverManager*** tries to load the driver through the current ***CLASSPATH*** *given in the System Environment of the computing machine***.** The ***DriverManager*** class uses the following piece of Java program to locate, load and link the named class.
Class.forName(driver).newInstance().
In case of ***oracle.jdbc.driver.OracleDriver*,** the driver class name can be located by
Class.forName(oracle.jdbc.driver.OracleDriver);
Now use the ***DriverManager*** class method ***registerDriver()*** to register the ***oracle.jdbc.driver.OracleDriver*** driver's class instance as:
DriverManager.registerDriver (new oracle.jdbc.driver.OracleDriver());
When above statement is executed, a new instance of the driver class is registered. It will not verify whether the connection is established or not. In order to establish the connection to the database, the following method of the ***DriverManager*** class is used

// define the Connection instance
***Connection sqlconn = null*;** // initially it is null
sqlconn = DriverManager.getConnection (url, User, Password);
Where url, User, and Password are declared as:

String url = "jdbc:oracle:thin:@dbserver:1521:infs";
String User = "user-name";
String Password = "user-password";
In the ***url*** string: ***dbserver*** is the name of the Oracle Server, ***1521*** is port of the machine on which this server is running and ***infs*** the instance of the **Oracle Database**. Other string variables are self -explanatory. When the connection is established and validated, the ***Application layer*** can be approached. A list of driver class names for different database management systems is given in Table 3.

In the above table, the name of the driver is also a driver class name, for example, for Oracle database system, the driver name ***Oracle.jdbc.driver.OracleDriver*** is also the driver class name, any instance of this class can be defined as **new** ***Oracle.jdbc.driver.OracleDriver()*** using constructor of this class [6,7,8], for example
Driver Driver_Name = new Oracle.jdbc.driver.OracleDriver();
The above statement creates a new instance of class ***Driver*** which can be registered with ***DriverManager*** class using ***registerDriver()*** method as
DriverManager.registerDriver(Driver_Name);

5.2 Application Layer

Application Interface: In Java programming language [8, 9], the application interface provides a means of using a general type to indicate a specific class. Three main application layer interfaces are ***Connection***, ***Statemen****t* and ***ResultSet*** classes. Each one of them is described below:

5.2.1 The Connection Interface:

A Connection ***object is obtained by using the*** DriverManager.getConnection() ***method call as***
Connection sqlconn = DriverManager.getConnection (url, User, Password);
where ***sqlconn*** is the ***Connection*** object returned by the called method ***DriverManager.get Connection (url, User, Password);***
where ***getConnection (url, User, Password)*** method uses three arguments ***url***, ***User*** and ***Password*** as described above.
Typical database connection include the ability to control changes made to the actual database stored through transactions [6,7,9]. When connection is created, it is in an ***auto-commit*** mode, that is, there is no **rollback** possible. After the connection from the driver is established, the application developer can set auto-commit to **false** by using ***setAutoCommit (boolean b)*** method. After setting this method call, the ***Connection*** will support both ***Connection.Commit()*** and ***Connection.rollback()*** method calls.
a-The ***Connection*** Class interface
The ***Connection*** class interface has the following methods:

i- Statement createStatement() throws SQLException : The ***Connection*** object will return an object of a ***Statemen****t* implementation such as
Statement sqlStatement = sqlconn.createStatement(); // use *sqlconn* Connection instance to create a statement
The ***Statement*** class object ***sqlStatement*** is implemented to execute a query if required and get a single ***ResulSet*** object

ii- PreparedStatement preparedStatement (String sql) throws SQLException: The ***Connection*** object implementation will return an instance of ***PreparedStatement*** object which is configured with **sql** string passed [8, 9]. The driver may then send the statement to the database if the driver handles the precompiled statements; otherwise the driver may wait until the ***PreparedStatement*** is executed by an ***execute()*** method

iii. void setAutoCommit (Boolean b) throws SQLException*:* This method sets a flag in the driver implementation that enables commit/rollback (false) or make all transactions commit immediately (true) as
sqlconn.setAutoCommit(false); // rollback all transactions

iv-void commit() throws SQLException: Makes all changes made since the beginning of the current transaction.

v- CallableStatement preparedCall(String sql) throws SQLException*:* The **Connection** object implementation will return an instance of a **CallableStatement. CallableStatements** are optimized for handling stored procedures. The driver may then send the **sql** string immediately when **prepareCall**() method is complete or may wait until an **execute** method executes.

vi- void rollback() throws SQLException*:* Drop all changes made since the beginning of the current transaction.
Mainly the ***Connection*** object interface is used to create a ***Statement*** object as
Connection sqlconn **;** **//** declare Connection object

Statement sqlStatement **; //** declare Statement object
sqlconn = DriverManager.getConnection (url, User, Password); **//** establish connection to the //database
sqlStatement = sqlconn.createStatement(); **//** create a statement object, to be used for executing a query sent to the database server

5.2.2 The Statement Interface: *Statement* Class methods

This interface is used to send **SQL** statements (**insert**, **delete**, **update** , **selec**t) to the database on the server and constructing corresponding result sets. This can also be used to create or drop tables from the database. ***SQLException*** is thrown if there is a problem with the connection of the database. The following methods are available with this interface [8,9].

i- ResultSet executeQuery(String sql) throws SQLException*:* Executes a single **SQL** query and returns the results in an object of type ***ResultSet***. This method can be used as
ResultSet rset ; // declare an object of a ResultSet
String sqlQuery = " select * from Dept "; **//** Dept is the name of the table on the database server
rset = sqlStatement . executeQuery(sqlQuery); **//** a result set is created

ii- int executeUpdate(String sql) throws SQLException : This method executes a single **SQL** query to return the number of rows affected rather than a set of results.

iii- boolean execute(String sql) throws SQLException: This method can be used in the following way:

a. To execute **SQL** statements that returns multiple result sets.
b. To execute for updating counts.
c. To execute stored procedures that return ***out*** and ***inout*** parameters.

This method is less commonly used in database processing than ***executeQuery()*** and ***executeUpdate()*** methods. The methods ***getResultSet(), getUpdate()*** and ***getMoreResultSet()*** are used to retrieve the returned data [7,8,9].

5.2.3. ResultSet Class Interface

The ***ResultSet*** interface defines the methods for accessing tables of data generated as a result of executing a ***Statement*** [5,6,7,8,9]. ***ResultSet*** column values may be accesses in any order, that is, they are indexed and may be selected by either the name or the number of the column. ***ResultSet*** maintains the current position of the row, starting first row of the data returned. The next() method moves to the next row of the data. The following program segment explains ***next()*** method.

```
ResultSet rset ;  // declare an object of a ResultSet
String sqlQuery = " select * from Dept ";              // Dept is the name of the table on the
database
                                                        //server
rset =  sqlStatement . executeQuery( sqlQuery);     // a result set is created

// processing of the resultset

if (rset.next()) {

    // processing statements goes here
}
```

The details of the ResultSet interface are discussed in [5,6,7,8,9].

6. APPLICATION INTERFACE-STRUCTURE CHART

Application interface is defined in Appendix A, Fig. 7.
It depends on the programmer which threads he /she wants to run to create a concurrency, for example, Thread-1 and Thread-2 can be executed in concurrent states to **insert** and **display** data at the same time. Concurrency programming is a tricky job. Similarly, Thread-1 and Thread-3 can be used concurrently to **update** and **display** data in the database. Different combinations of these threads can be used to compromise between the execution and the complexity of the code developed for the application. In a single thread execution, activities take place in sequential order [5,6,7,8,9]. The complete program is given in the following section. This program is used to retrieve and display data under thread-1. The development tool used is **NetBeans** IDE 3.5.1. This IDE has partially built-in Java programming document, which can be used to code the program, thereby, minimizing the development time for business applications.

7. DESCRIPTION OF THE PROGRAM USING A SINGLE THREAD

This program is defined as a single class ***jdbcDbRetrieval*** which extends to a class ***JApplet*** and is running under a ***thread*** control to create a concurrency or parallelism. This program is running under Windows Operating System to check the effect of multithreading techniques built into Java Programming Language [6,7,8,9]. This is a unique program in itself. The other database functions such as Insert, Update and Delete are also programmed but are not given in this paper. Each one of them is an applet running under a single thread and coordinating the other threads when required.

8. CONCLUSION

Java Programming Language can be used to development Distributed or Concurrent business applications in order to decrease the development time and other resources. Java API is an important part of the application development stage where a large number of built-in class and their methods are available to take full advantage of Java Development Kit. It also provides a guideline to those who are interested in developing business applications which can be run in parallel. Using Java applications are implemented and installed on different platforms with little or no change in the coding of the applications. To incorporate all these concepts and tools a complete program to implement retrieval operation of the database is given in Appendix A.

REFERENCES

[1] R. Greg (2001): Principles of Database Systems with Internet and Java Applications, Addison Wesley New York.

[2] Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden (2005): Modern Database Management, Seventh Edition, Pearson-Prentice Hall, U.K.

[3] R. Peter; C. Carlos (2002): Database Systems, Fifth Edition, Course Technology, Thomson Learning, U. K.

[4] V. Michael (2004): Database Design, Application Development and Administration, Second Edition, McGrawHill, Toronto, Canada.

[5] H. M. Deitel; P. J. Deitel (2002): Java: How to Program, Fourth Edition, Prentice Hall, New Jersey, U. S. A.

[6] H. M. Deitel; P. J. Deitel (2003): Java: How to Program, Fifth Edition, Prentice Hall, New Jersey, U. S. A.

[7] H. M. Deitel; P. J. Deitel (2005): Java: How to Program, Sixth Edition, Prentice Hall, New Jersey, U. S. A.

[8] B. Kurniawan (2002): Java for the Servlets, JSP, and EJB, Techmedia, Delhi, India.

[9] H. M. Deitel; P. J. Deitel; S. E. Santry (2002): Advanced Java 2 Platform: How to Program, Prentice Hall, New Jersey, U. S. A.

[10] D. Cohoon (2004): Java 1.5: Program design, McGrawHill, U. K.

[11] C. Thomas Wu (2004): An Introduction to Object-Oriented Programming with Java, Third Edition, McGraw-Hill, U. K.

[12] J. Adolph Palinski (2003): Oracle 9i Developer: Developing Web Applications with Forms Builder, Thomson, U.K.

[13] J. Morrison; M. Morrison (2003): Guide to Oracle 9i, Thomson, U. K.

[14] M. A. Ajiz (2002): E-Commerce Systems development: Case Study, Pakistan Journal of Applied Sciences 2 (2): pp.245-259, Lahore, Pakistan.

[15] R. Greenlaw; E. Hepp (1999): Fundamentals of the Internet and World Wide Web, McGraw-Hill, Toronto, Canada

Appendix A

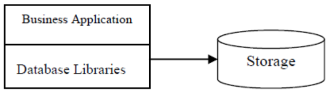


Figure 1. Single-tier database design

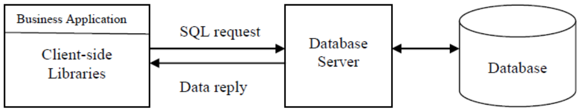


Figure 2. Two-tier database design

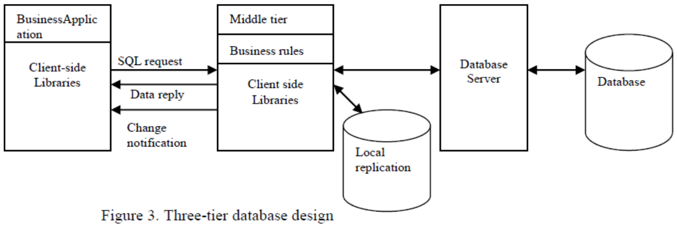


Figure 3. Three-tier database design

Figure 6: Multi-tier JDBC database design

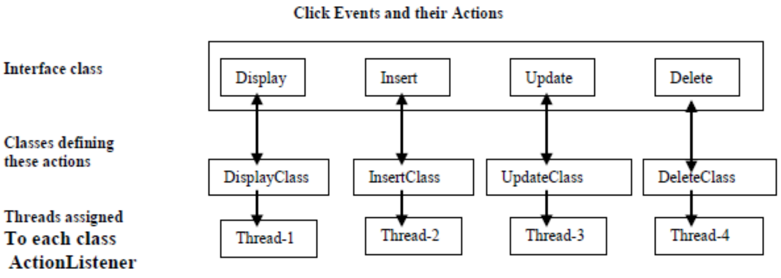


Figure 7: Application Interface

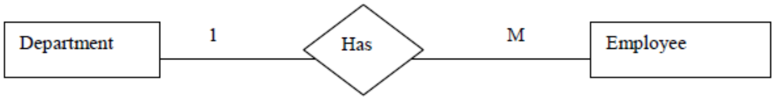


Fig. 4: One-to-many relationship

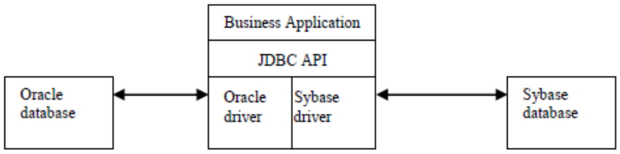
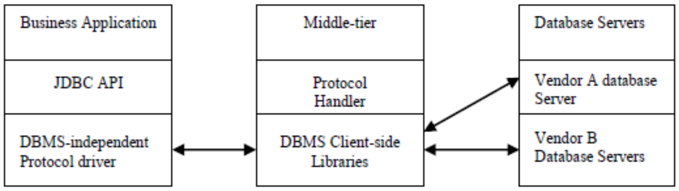
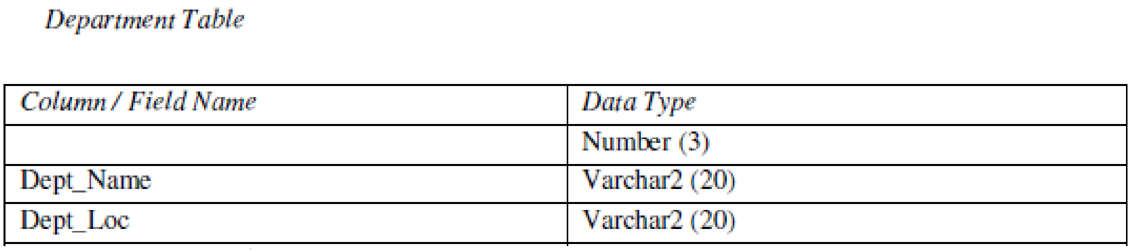


Figure 5: Single-tier JDBC database design strategy

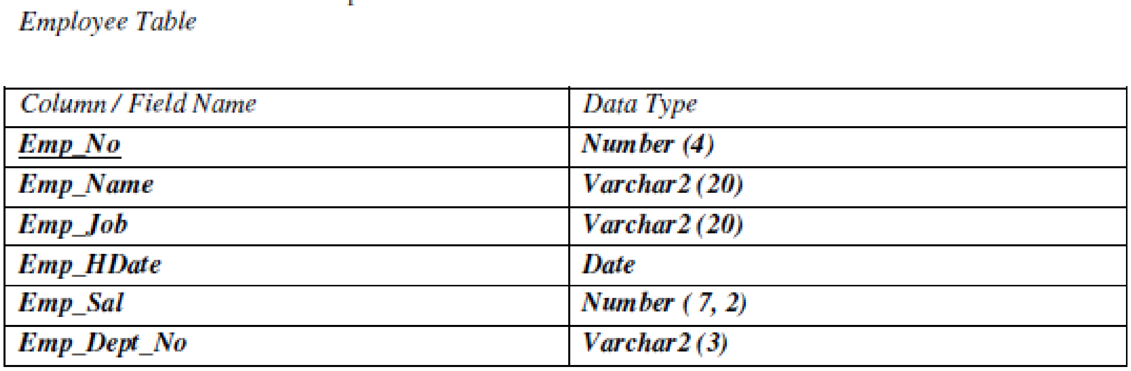




Department Table

Column / Field Name	*Data Type*
	Number (3)
Dept_Name	Varchar2 (20)
Dept_Loc	Varchar2 (20)

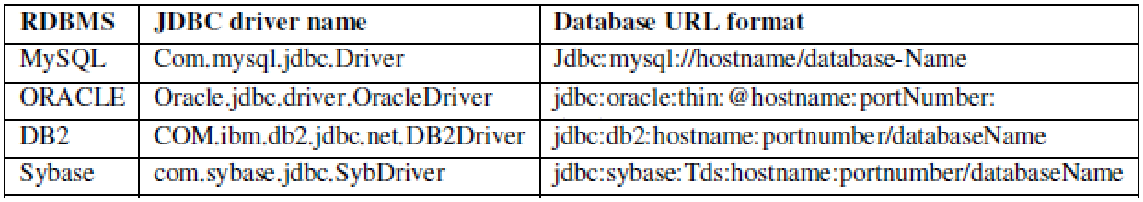
Table 1: Department



Employee Table

Column / Field Name	*Data Type*
Emp_No	***Number (4)***
Emp_Name	***Varchar2 (20)***
Emp_Job	***Varchar2 (20)***
Emp_HDate	***Date***
Emp_Sal	***Number (7, 2)***
Emp_Dept_No	***Varchar2 (3)***

Table 2: Employee



RDBMS	**JDBC driver name**	**Database URL format**
MySQL	Com.mysql.jdbc.Driver	Jdbc:mysql://hostname/database-Name
ORACLE	Oracle.jdbc.driver.OracleDriver	jdbc:oracle:thin:@hostname:portNumber:
DB2	COM.ibm.db2.jdbc.net.DB2Driver	jdbc:db2:hostname:portnumber/databaseName
Sybase	com.sybase.jdbc.SybDriver	jdbc:sybase:Tds:hostname:portnumber/databaseName

Table 3 : List of JDBC Drivers and their Classes

AUTHORS

Dr. Raied Salman received his second Ph.D. in computer science from the Department of Computer Science at Virginia Commonwealth University (Richmond / USA). He also received his first Ph.D. from Brunel University (England / UK) in Electrical Engineering and both Bachelor degree and Master degree of Electrical Engineering from The University of Technology (Baghdad / Iraq). His research interests include machine learning and data mining.

INTENTIONAL BLANK

CRITICAL SUCCESS FACTORS (CSFS) OF ENTERPRISE RESOURCE PLANNING (ERP) SYSTEM IMPLEMENTATION IN HIGHER EDUCATION INSTITUTIONS (HEIS): CONCEPTS AND LITERATURE REVIEW

Ashwaq AlQashami[1] and Heba Mohammad[2]

[1]Al-Imam Muhammad Ibn Saud University, College of Computer and Information Sciences, Information Systems Department,Riyadh, Saudi Arabia
asqashami@sm.imamu.edu.sa
[2]Al-Imam Muhammad Ibn Saud University, College of Computer and Information Sciences, Information Systems Department,Riyadh, Saudi Arabia
hkmohammad@ccis.imamu.edu.sa

ABSTRACT

Nowadays, Information Technology (IT) plays an important role in efficiency and effectiveness of the organizational performance. As an IT application, Enterprise Resource Planning (ERP) systems is considered one of the most important IT applications because it enables the organizations to connect and interact with its administrative units in order to manage data and organize internal procedures. Many institutions use ERP systems, most notably Higher Education Institutions (HEIs). However, many projects fail or exceed scheduling and budget constraints; the rate of failure in HEIs sector is higher than in other sectors. With HEIs' recent movement to implement ERP systems and the lack of research studies examining successful implementation in HEIs, this paper provides a critical literature review with a special focus on Saudi Arabia. Further, it defines Critical Success Factors (CSFs) contributing to the success of ERP implementation in HEIs. This paper is part of a larger research effort aiming to provide guidelines and useful findings that help HEIs to manage the challenges for ERP systems and define CSFs that will help practitioners to implement them in the Saudi context.

KEYWORDS

Enterprise Resource Planning (ERP) system, Critical Success Factors (CSFs), Higher Education Institutions (HEIs), ERP implementation, Higher Education

1. INTRODUCTION

Information Technology (IT) has brought the best products to enrich various aspects of modern life; no organization can be effective without the adoption of the latest available technology. An Enterprise Resource Planning (ERP) system is one of the technologies used for the best running of organizations to attain effectiveness and efficiency. This system has been defined by many researchers as an integrated information system (IS) and a comprehensive software package that

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 81–98, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51508

integrates and controls all the business processes and functions in an organization to institutionalize the sharing of organizational data resources [1, 2, 3, 4].

The ERP system consists of software support modules including utilities formarketing and sales, field service, product design and development, production and inventory control, procurement, distribution, industrial facilities management, process design and development, manufacturing, quality, human resources, finance and accounting, and information services [5]. It helps the different departments of an organization to move information among different processes, reduce costs, increase operational efficiencies, improve business process management, facilitate communication, share information and knowledge across organizational units, and improve decision making capability [6]. Because of these improved features and through the development of business and administrative procedures, many organizations around the world have implemented or updated their current management IS with an ERP system or are in the process of implementing such a system. Despite all the benefits of these systems, implementation is costly, time consuming and complicated, requiring large investments in the fields of planning, consulting and implementing software projects. About 60–90% of organizations fail in the implementation of the ERP system [5, 7, 8], while about 90% of ERP projects go over time or over budget [9, 10]. Accordingly, many challenges and obstacles may arise that hinder successful implementation and affect the benefits intended from the system.

The Higher Education Institutions (HEIs) sector is one of the most important sectors seeking to keep pace with technological development and to benefit from ERP systems to accelerate and simplify the management of data and internal procedures while reducing the cost and increasing the efficiency of institutional performance. Unlike business sectors, which seek commercial profit, HEIs belong to the non-profit governmental sector. HEIs spend huge amounts to move to these advanced systems, but many difficulties may arise. Such difficulties often end the process of ERP implementation, resulting in failure because HEIs –as part of the governmental sector– have a unique nature where administrative structures are inert and employees tend to resist the idea of change. Therefore, HEIs have an urgent need to concentrate on changing the processes prior to implementing the technology. Thus, the role of the top management support is vital for planning and implementing ERP systems, as it is necessary to stimulate the organization and employees before implementation and engage in effective communication with staff to increase the probability of success [11, 12].

Previous studies have suggested that ERP software system implementations is complicated and time consuming [5, 7, 10, 13, 14, 15]. In addition, these studies have pointed out that there is no sure method of achieving success in implementing ERP systems; many ERP implementation projects fail or go over time and over budget. It is noteworthy that several previous studies have focused on the implementation of ERP systems in the business sector [5, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24], but there is lack of studies on implementation in the HEI sector around the world [25, 26, 27, 28, 29]. Moreover, these studies have argued that the university community needs research attention to achieve more relevant knowledge regarding ERP implementation [26]. Furthermore, there is an urgent need to identify the success factors that lead to the successful implementation of these systems [28], since the failure rate in HEIs is higher than that in the business sector [11, 30]. Moreover, Saudi HEIs are in the early stage of technology development because of the lack of implementation of ERP systems in this country [30, 31]. Therefore, it is crucial to study critical success factors (CSFs) to reduce the failure rate of ERP systems in Saudi HEIs.

2. DEFINITION OF ERP SYSTEMS

Several definitions of ERP systems have been offered by a range of authors in the literature. For example, Gable [32] described ERP systems as comprehensive software packages that seek to integrate and automate the complete range of business processes and functions to present a holistic view of the business from a single information and IT architecture. Rosemann and Wiese defined ERP systems as “customizable, standard application software which includes integrated business solutions for the core processes (e.g. production planning and control, warehouse management) and the main administrative functions (e.g. accounting, human resource management) of an enterprise” [33, p. 1].

Previously, ERP systems were perceived as integrated software applications that control and manage different departmental functions such as inventory control, accounting, finance, and human resources (HR) in a single centralized system with a common database. Now, ERP II — the second generation of ERP systems— has been introduced with advanced features to deal with multiple business units, such as customer relationship management and supply chain management. Moreover, it integrates Internet-enabled applications for e-business, allowing access at anytime and from anywhere. The term ERP evolution is widespread and describes integrated information systems appropriate to any organization, regardless of geographic location and size [14, 34, 35].

Many researchers have provided broad definitions of ERP systems. However, a recent and comprehensive definition of ERP systems was provided by Beheshti as “a set of business applications or modules, which links various business units of an organization such as financial, accounting, manufacturing, and human resources into a tightly integrated single system with a common platform for flow of information across the entire business. With the use of the Internet as a business medium, organizations can use the expanded version of ERP, ERP II, to connect their internal business systems with the systems of customers and suppliers” [36, pp. 184–185].

Each ERP system contains many different modules referring to business functions such as HR; some business functions have more advanced and powerful modules than others. A typical ERP system may consist of the following software modules: HR Management, Accounting and Finance, Procurement Management, Manufacturing, Distribution and the Supply Chain [37]. Moreover, several vendors offer ERP systems in the marketplace; the top four vendors are SAP, Oracle, Baan, and PeopleSoft [21, 38].

3. MOTIVATIONS FOR ADOPTING ERP SYSTEMS AND THEIR BENEFITS

There are many reasons to adopt ERP systems. It is important for the organization to understand the reasons for deciding to adopt ERP systems so that they can take advantage of the full benefits.

The main reasons for implementing an ERP system can be summarized as follows: providing an integrated business computing solution, improving a company’s ability to compete in the marketplace, improving business processes and internal efficiency of workflow and reducing the overhead costs in an institution through computerization, enhancing the decision-making process by providing accurate and updated organization-wide information. Improvement in all of these areas will enhance company performance [36, 39, 40, 41, 42].

A comprehensive framework for assessing the benefits of ERP systems has been proposed by Shang and Seddon [43, 44]. This framework classifies the types of benefits that organizations can obtain by using ERP systems along five dimensions as follows:

- Strategic: Supporting business alliance and business growth, generating product differentiation, and building cost leadership, business innovations, and external linkages with customers and suppliers;
- Operational: Cost and cycle time reduction, quality, productivity and customer service improvement;
- Managerial: Better resource management, improved decision making and planning, and performance improvement;
- IT infrastructure: Building business flexibility for current and future changes, IT cost reduction, and increased IT infrastructural capability; and
- Organizational: Supporting organizational changes, and facilitating business learning, empowerment, and building a common vision.

Numerous studies [40, 41, 43] have listed the most important attributes of ERP systems and their ability to effectively improve business'organizational processes, including the following:

- Automating, coordinating, and integrating business processes across organizational locations and functions.
- Sharing common data and practices across the whole enterprise to reduce errors.
- Producing, accessing, and managing information in a real-time environment available anywhere and anytime to facilitate rapid and better decision making and cost reductions.
- Providing a user-friendly web interface system to corroborate interactivity, as such an interface can improve integrated portals for an extensive variety of administrative functionalities.
- Enabling effective and useful conduct of a new business process, such ase-government, e-learning, e-commerce, e-procurement and e-portfolio.

4. Challenges To ERP Implementation

Although implementing an ERP system has significant benefits, doing so successfully is a challenge. ERP systems are highly complex and require a comprehensive risk strategy; moreover, they are very costly and have a high failure rate even under ideal circumstances [13]. They often require long implementation times and significant resources [5, 7, 10, 13, 14, 15, 45]. According to Zhang et al. [15], on average, ERP projects were 178% over budget, took 2.5 times as long as projected, and delivered only 30% of the planned benefits. In addition, many barriers appear that affect successful implementation, including substantial organizational problems such as employee resistance to change [13, 24, 46]. Thus, the decision to implement an ERP system is a difficult undertaking for any organization.

Over the years, many companies have implemented ERP systems, but many others have faced implementation failure or ended up going over budget and experiencing delays [14, 38, 47, 48, 49, 50]. Approximately 90% of ERP projects end up late or over budget [9, 10, 13]. Markus and Tanis [51] defined success from the viewpoint of managers and implementation consultants as completing the ERP project implementation on time and within budget. Therefore, chief executive officers and senior executive teams must be deeply involved and have a strong commitment to the ERP project to achieve successful implementation [38].

An understanding of the main reasons why many ERP implementation projects have failed could be a recipe for success in a new project. Umble et al. [24] summarized reasons for project failure into 10 categories. They contended that ERP could fail because of a lack of clearly defined strategic goals, lack of commitment of top management, poor project management, resistance to change in the organization, poor selection of an implementation team, lack of data accuracy, inadequate education and training such that users cannot run the ERP system, lack of adaptation of performance measures to ensure that the organization changes, lack of resolution of multi-site issues, and technical complications.

5. SUCCESS FACTORS FOR ERP IMPLEMENTATION

The concept of CSFs was developed in the early 1960s. According to Rockart [52], Ronald Daniel first discussed the idea of CSFs in the management literature, stating that information analysis must focus on “success factors” when as a new approach to help achieve organizational goals.

Rockart [52] developed the idea of identifying the CSFs from the viewpoint of chief executives, pointing out that the process of identifying CSFs helps to ensure that these factors receive the necessary attention and are carefully managed by an organization. Rockart defined CSFs as “the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization”[52, p.85]. Rockart and Bullen affirmed that “CSFs are the few key areas where ‘things must go right’ for the business to flourish and for the manager’s goals to be attained” [53, p.7]. Meanwhile, Pinto and Slevin described CSFs as “factors which, if addressed, significantly improve project implementation chances” [54, p. 22].

In the ERP context, Rabaa’I defined CSFs as “a set of activities that needs special considerations and continual attention for planning and implementing an ERP system” [28, p. 137]. They have also been defined as “factors needed to ensure a successful ERP project“ [17, p. 31]. Thus, CSFs are particularly useful, as they provide clear insight and guidance on where to focus special consideration and resources and continual attention in planning for successful ERP project implementation [55].

It is important to improve the implementation success of ERP systems, identify the CSF, and understand the critical factors that constitute successful ERP implementation at each phase [56]. Thus, the critical factors involved in ERP implementation can be given more, and proactive approaches can be developed to counter the high failure rate of ERP implementation [21, 52]. Ultimately, this will enhance the likelihood of achieving higher success levels, cost savings, time savings, quality and efficiency in their ERP system [57].

6. ERP SYSTEM IN HEIs

Instead of developing an IT system in-house, many local government organizations are turning to commercial off-the-shelf (COTS) ERP systems solutions offered by commercial vendors that support core administrative processes such as budgeting, accounting, procurement, performance, and HR management by integrating the data required for these processes in a single database. This allows them to plan their IT resources more effectively, manage their data and legacy systems, and increase the efficiency of the institutional performance [58]. Recently, ERP systems have been applied to new institutional contexts (i.e., financial services, public sector, healthcare, and higher education). Therefore, the HEI sector as one of the main parts of government has been strongly influenced by global trends to adopt new technologies. HEIs begin implementing ERP systems and replace their old systems to overcome the limitations of legacy systems, support all business functions (administrative, accounting, organizational, etc.), improve their management and administration systems, manage their operations and make them more transparent, and achieve performance improvements.

The investment in ERP systems implementation represents the biggest investment in information and communications technology (ICT) for HEIs [50, 59]. It is tempting to see HEIs as unique institutions [50]; however, according to Lockwood [60], there are many similarities and differences between HEIs and business organizations, so that the HEIs face many problems common to most modern business corporations, including coordinating resources, stimulating and facilitating the enterprise among staff, and controlling costs. Meanwhile, the uniqueness of HEIs is based on a combination of different characteristics, namely the complexity of purpose, limited measurability of outputs, autonomy and dependency from wider society, a diffuse structure of authority and internal fragmentation. In addition, HEIs are fundamentally different from business organizations due to their unique decision-making processes, where each executive member is capable of independent decision-making and behavior [50, 61]. Further, HEIs are generally more resistant to change than business corporations due to the loosely coupled and autonomously operating administrative and academic units [25]. Furthermore, Bologa et al. [62] pointed out that the communication in HEIs is more difficult than in companies due to the large number of very different groups with different interests and objectives in different fields; thus, there are no clear formal communication structures in HEIs. The characteristics that make HEIs different require a distinct project management approach.

The main advantages of implementing an ERP system in HEIs are as follows: lower business risks; improved services for the faculty, students, and employees; increased income and decreased expenses due to improved efficiency; and improved information access for planning, organizing, and managing the institution where different departments share an integrated database [63]. Furthermore, Sabau et al. [64] introduced many ERP benefits for universitiesin terms of business and technical viewpoints.

The *business benefits* include:

- Campus-wide integration of a conventional system.
- Enhanced internal communications.
- Reduction or elimination of manual processes.
- Improved strategic decision making and planning capabilities.
- Self-service environment for students and faculty.

- Higher availability of administrative systems.
- Support of sophisticated and advanced data analyses for use in decision making.
- Integrated workflow for the industry's best practices and decreased dependence on paper.

The *technical benefits* include:

- Decrease or eliminated need for backup systems.
- Platform for re-engineering business practices and continued process enhancements.
- Maintenance and development of consistent data definitions.
- Accessible, user-friendly administrative and student support services.
- Increased data integrity, reliability, and validity.
- Guaranteed system-wide security and protection of confidential information.
- More seamless integration between technology and education delivery by providing a single platform based on new technologies and access to data in real time.

Nevertheless, a limited number of integral ERP solutions have been implemented in HEIs, and there is a need to introduce such ERP systems. Still, the risks related to the implementation of ERP systems in HEIs are relatively high due to the high degree of complexity [59]. According to Pollock and Cornford, ERP systems are "refashioning the identity of universities and accompanied by tensions in which ever site it is implemented" [50, p. 32]. Therefore, the implementation of these systems in the HEI sector is raising new organizational issues [65]. These systems were initially designed for the corporate sector, with little effort to make them fit universities' business processes; consequently, they either adjust the universities business processes to fit the ERP system or customize the ERP system to fit the universities' business processes [66]. In addition, Davison [67] identified cultural differences from that of North America through a case study on a Hong Kong University's implementation of an ERP system, including beliefs concerning providing access to information and miscommunication and difficulties in reengineering organizational processes. A number of researches have shown a high failure rate in the implementation of ERP systems; for instance, Cleveland State University (1998) took legal action against the ERP vendor when they found their new system could handle only half of their transaction volume. The university continued with the implementation of ERP despite rising costs, with a final cost of $15 million, which exceeded the initial forecast by $10.8 million. Likewise, ERP implementation costs for Ohio State University rose from an initial planned amount of $53million to $85 million. The University of Minnesota had a comparable experience when planned projected costs of $38 million finally reached $60 million [68]. This illustrates the importance of minimizing ERP implementation failure in the HEI sector.

7. RELATED WORK

ERP systems have received substantial consideration in both academia and practice. Numerous research articles about ERP systems have been published, covering several topics and issues. Furthermore, a number of ERP literature reviews have been conducted [14, 26, 38, 45, 57, 69, 70, 71, 72, 73] that provide an overview of existing ERP literature from varied perspectives.

Ngai et al. [14] reviewed 48 articles to illustrate the disparity betweenthe 10 different countries/regions surveyed and the recommended empirical evidence for criticality of the 18 identified ERP success factors. The study results showed the most frequently cited critical factors for successful implementation of ERP systems are top management support, as well astraining and education. Moreover, a clear and defined project plan was another frequently cited factor in all the countries and regions.

In addition, Gargeya and Brady [38] identified six groups of factors leading to success in ERP implementation by reviewing published articles that reported SAP implementations in 44 companies from different industries based on conducting content analysis. The primary factors for successful implementation of SAP ERP projects include working with SAP functionality and maintaining scope and the cooperation of the project team/management support/consultants. Other success factors are the internal readiness/training, adequate system testing, organizational diversity, and planning/development/budgeting.

A comprehensive taxonomy of CSFs for ERP system implementation was presented by Dezdar and Sulaiman [45]. They analyzed the content of 95 articles publishedbetween1999 and 2008; then, they arranged 17 identified ERP success factors using the frequency count method. They found that the most frequent CSFs include top management support and commitment, project management and evaluation, business process reengineering and minimum customization, ERP team composition, competence and compensation, and the change management program.

However, Finney and Corbett [57] recommended 26 CSFs based on the results of a comprehensive compilation and analysis of ERP implementation success factors, which they identified using content analysis and grouped into strategic and tactical categories. The study reveals that the five most significantly cited CSFs are top management commitment and support, change management, business process reengineering (BPR) and software configuration, training and job redesign, and the best and brightest project team. However, change management has emerged as the most widely cited CSF.

Since ERP literature is a wide topic, we centered our review on ERP implementation in HEIs, which provide a more detailed analysis and deeper understanding of CSFs in ERP implementation within this sector.Abugabah and Sanzogni [11] collected critical literature reviews of the implications of ERP systems in HEIs, especially in Australia. The study shows that system users at all levels play a major role in defining the feasibility of ERP implementation. The researchers discuss the importance of establishing the requisite criteria to evaluate the performance of the system through the performance of its usersafter providing necessary training. It is important not to neglect other characteristics, such as technical and managerial aspects. The researchers stress the importance of aspects that influence the performance of staff, the quality of services provided, and the output of the system. Therefore, they try to gather the evaluation of ERP implementation while considering user and organization perspectives.

Furthermore, a study by Rabaa'i [28] concentrates on ERP implementation and evaluation, also in Australian HEIs. He discussed the CSFs identified by previous studies and their importance to answer the following research question: What are the key critical factors for ERP implementation success in a university environment? Based on a literature review, 12 CSFs for ERP implementation were identified using the frequency analysis. Change management and top management commitment and support are the most widely cited CSFs. Other factors include project management, business process reengineering and system customization, user training, cross-functional implementation teams, visioning and planning, consultant selection and relationship, an effective communication plan, ERP system selection, ERP systems integration, and post-implementation evaluation measures.

8. RESEARCH METHODOLOGY

The researchers conducted an extensive literature review, analyzing more than 50 articles published over a period of 13 years (2002–2015) to address the following question: ' *What are the most imperative critical factors for ERP implementation success?'* In this study, articles from journals, book chapters, conference proceedings, and dissertations were identified, analyzed, and classified. These articles were identified through a computer search of Management Information System (MIS) journals and number of databases including Emerald, ScienceDirect, Proquest Computing, IEEE/Xplore, EBSCOhost, SpringerLink, ACM Digital Library and Google Scholar.

The articles were selected was based on the following search terms and keywords: “enterprise resource planning success factors,”“ERP implementation success,”“ERP implementation success in higher education institutions,”“critical success factors for ERP implementation in higher education,”“critical success factors for enterprise systems,”“enterprise resource planning successful implementation,”“success factors of enterprise systems” and “CSFs of ERP system implementation in HEIs.”

9. SUCCESS FACTORS FOR ERP IMPLEMENTATION IN HEIS

As mentioned above, an ERP redefines business operations and plays an important role in managing business processes in many organizations. In addition, there are many previously identified factors that could influence the successful implementation of ERP systems. Recently, universities have implemented ERP systems, but there is a lack of research focusing on the implementation of ERP systems in the HEI sector around the world [26, 29, 74]. Furthermore, some studies have identified the urgent need to identify success factors that lead to the successful implementation of the system, as failure rate of ERP systems in HEIs is higher than that in the business sector [11, 30].

Allen et al. [75] investigated the issues associated with ERP implementation via four in-depth case studies of HEIs in the UK that were in the process of implementing ERP systems to investigate whether the systems provided a feasible IS strategy for HEIs through interviews and reviews of secondary documentation. They adopted Pinto and Slevin's [54] and Holland et al.’s [76] CSF models, which include strategic issues and specify the need for a project mission, top management support, and project schedule outlining individual action steps for project implementation. Clearly, these issues are most important during the rollout of a project. Meanwhile, tactical elements such as communicating with all affected parties, recruiting the necessary technical and business specialists for the project team, obtaining the necessary underlying technology, user acceptance, and monitoring and feedback at each stage gain their importance in the implementation phase.

Chatfield [77] investigated the implementation factors that affect ERP system success in universities,including a decrease in implementation costs. The findings suggest that ERP system quality is improved by effectively training users, changing management, and providing support strategies for the users and the organization’s culture. This requires involving university personnel in the implementation to reduce costs and making use of the personnel's existing knowledge of the organization.

Seo [12] in her thesis, focused on the challenges of ERP implementation in corporate and university environmentsby conducting two case studies tocompare the similarities and differences, specifically between the Massachusetts Institute of Technology (MIT) and the Engineering Company (ENGCO). This research presents the top 12most frequently cited CSFs from previous studies. After evaluating these factors, the following CSFs were proven effective in the implementation of ERP in the university environment: change management, communication planning, ERP systems integration, andtop management commitment and support. Moreover, Seo found that ERP failure results from the unique nature and decision support methodology of the university as well as the limited flexibility of university systems. The researcher also mentioned that university administration support in achieving the benefits of ERP systems could be the cornerstone for success, along with other success standards, such as team composition and strong communication across the organization.

Olugbara et al. [78] identified, validated, ranked and classified ERP success factors with reference to HEIs and described expert assessments to validate the relevance of the identified ERP success factors in the educational setting. Moreover, they used principal component analysis to reduce the dimensions and rank ERP success factors, as well as cross-impact analysis to classify ERP success factors. The study identified the following 10 CSFs influencing the effective implementation of ERP systems in African HEIs: top management support, management of expectations, business process reengineering, project team composition and competence, education and training of users, interdepartmental cooperation and communication, involvement of users in systems development and integration, culture of resistance within an organization, vendor and consultant support for users, and system changes and upgrades to new versions.

Somers and Nelson [79] developed a unified CSF model for industries in United States that described the importance of 22 CSFs identified across the stages of ERP implementation through responses from 86 organizations within variety of industries, including the education sector. The research sample targeted in the interviews and questionnaire consisted of senior-level IS executives. The authors identified and ranked the top five CSFs for ERP implementation as top management support, project team competence, inter-departmental cooperation, clear goals and objectives, and project management. Furthermore, the study suggested that the most important factor for executives is top management support.

Finally, Nah and Delgado [80] reviewed the literature to develop a comprehensive list of CSFs related to ERP implementation and upgrades. The researchers conducted two case studies at a university and a public company using interviews and questionnaires to collect data. They then organized the results into seven main categories. The results showed that the main factors involved in the successful implementation across the four phases of the ERP lifecycle are as follows: (1) a business plan and vision; (2) change management; (3) communication; (4) ERP team composition, skills, and compensation; (5) project management; (6) top management support and championship; and (7) system analysis, selection, and technical implementation.

10. ERP IMPLEMENTATION WITH SAUDI HEIs

With Saudi Arabia's orientation toward implementing e-government, whose significant benefits are attributable to the institutions and national economy of the Kingdom of Saudi Arabia, and due to government pressure to improve operational efficiency within their institutions, an e-government program (Yesser) was established that enables the implementation of e-government and raising the public sector's productivity and efficiency [81]. Saudi HEIs began implementing ERP systems to increase their efficiency and to automate their administrative procedures. As major Saudi government institutions, HEIs have been strongly influenced by the government trend to adopt new technologies.

A few research studies on the implementation of the ERP system in Saudi HEIs have concentrated on general technical and users' perspectives [7, 30, 82]; these have been limited to study the experience of King Saud University (KSU) in ERP implementation. It should be noted that the ERP system of KSU is called MADAR; it is not an ERP global software system such as SAP, Microsoft Dynamics, or Oracle E-Business Suite (EBS); rather, it has been locally designed due to budget constraints, a shortage of skilled users, and the greater flexibility for customization with the governmental policies [83]. Therefore, the success factors and obstacles may differ from the experiences of other universities that have implemented global ERP software systems.

Al-Hudhaif [82] investigated the factors affecting the ERP implementation from the user's perspective at KSU. A theoretical framework was developed and four hypotheses were explained to look at the status of system implementation at this university. The results showed the total success is dependent on the user satisfaction. In addition, the researcher found a significant relationship between satisfaction level and challenges to implementation. He suggested that the commitment of university head administration in ERP adaptation plays the major role of success implementation. However, the study suggests no relationships between the training factor and success of ERP implementation.

Moreover, Aldayel et al. [30] conducted a case study for the CSFs of ERP implementation in higher education from technical and user perspective. Their case study had been conducted at KSU which implemented MADAR system. The results showed that the most important CSFs in ERP implementation from a technical point of view were project management and ERP system selection. Other factors included stakeholder participation, business process reengineering and customization, top management commitment and support, ERP team composition, ERP systems integration, choice of supplier and its support, scope of implementation, and consultant participation. From the user's perspective, the most important factor was training.

11. CSFs FOR ERP IMPLEMENTATION IN HEIs

The results of the abovementioned research studies on the success factors of ERP implementation describe the problem complexity using a variety of approaches. The CSFs and the results of the research studies differed substantially, showing that the factors leading to success are complex and cannot occur in isolation. Indeed, they overlap and are hard to separate [38]. Some researchers have suggested that the CSFs of ERP implementation identified in the literature in terms of private sector organizations are equally applicable to organizations in the public sector. Furthermore, they pointed to additional CSFs in public sector organizations, where it is more difficult to successfully achieve ERP implementation. The government acquisition rules must be revised to align with the CSFs [38, 84, 85]; thus, CSFs need to be identified and adapted to the public sector.

The aim of this paper is to define the main important factors contributing to the success of ERP implementation in HEIs. Table 1 shows the 13 factors identified as critical to ERP implementation success from previous studies conducted in the HEIs sector.

Table 1. The most important ERP CSFs extracted from the literature.

#	Factors	References	Definition
1	**Top management commitment and support**	[25, 28, 29, 30, 62, 64, 66, 75, 78, 79, 80]	There is enough support from senior management in the ERP project. Top management must be willing to be involved and commit to allocating valuable resources to the implementation effort.
2	**Change management**	[25, 28, 29, 30, 62, 64, 66, 75, 78, 79, 80]	A primary strategy, strong institutional identity and structured approach are needed to create a comprehensive environment to ensure the successful implementation and smooth transitioning to the ERP system.
3	**Project management**	[25, 28, 29, 30, 62, 64, 66, 75, 78, 79, 80]	Effective management of the ERP project, including defining the project scope, goals, objectives, schedule and strategy and careful tracking of ERP project progress to plan, coordinate and monitor various defined activities in different stages of ERP implementation.
4	**Project champion**	[62, 66, 79, 80]	A project leader who plays a critical role in the implementation of ERP and makes the project work by setting goals and effecting legitimate changes.
5	**System customization**	[25, 28, 29, 30, 66, 79, 80]	Modification of the ERP package according to the institution's needs to fit its existing business process.
6	**Business process reengineering (BPR)**	[25, 28, 29, 30,62, 66, 75, 78,79, 80]	Changes in the work process that arise with ERP system implement on to fit and adapt the functionality of the system package instead of trying to modify the ERP system to fit the organization's current business processes.
7	**ERP implementation team**	[25, 28, 30, 62, 64, 66, 78, 79,80]	The team should consist of the best, most skilled people in the institution. Necessary capabilities include team leadership, cross-functional team representation from all business units and strong commitment to the implementation duties.

#	Factors	References	Definition
8	**Consultant selection and relationship**	[28, 30, 62, 66, 75, 78, 79, 80]	The extent to which ERP consultants who are expert and knowledgeable about the installation are part of the implementation process. It is important to arrange for knowledge transfer from the consultant to the project manager and staff.
9	**Effective communication plan**	[25, 28, 29, 62, 64, 66, 75, 78, 79, 80]	Covering and sharing information, scope, activities and objectives between the ERP project team members and communication of the results and goals at each ERP implementation stage to the rest of the institution.
10	**Active partnership with vendor**	[25, 30, 66, 75, 78, 79, 80]	Support ranging from technical assistance to training that can reduce the cost of implementation; the organization cam gain other benefits from partnerships with the vendor and use the vendor's customization tools.
11	**ERP system selection**	[28, 29, 30, 64, 66, 74, 78, 79, 80]	Careful ERP software package selection that matches the organizational needs, business processes and practices.
12	**System integration**	[28, 30, 80]	Good integration between the ERP system and other systems in the institution to smoothly share and transfer information.
13	**Post-implementation evaluation and management**	[28, 66]	All projects require some kind of post-evaluation through the exchange of information between the project manager and project team members and analysis of user feedback.

12. CONCLUSION

This paper gave a general overview of the implementation of ERP systems, including definitions, motivations for adopting ERP systems, and their challenges. Furthermore, it explained the CSFs concept and provided a comprehensive overview of the literature on ERP implementation success factors.

This research reviewed the previous literature on ERP implementation with a focus on success factors for ERP system implementation in the HEI sector worldwide and in Saudi Arabia. The aim of this was to fill the gap in research regarding the implementation of the ERP system in this sector, particularly since the failure rate of ERP systems in HEIs is higher than that in the business sector. This paper identified and defined the 13 most important ERP CSFs extracted from previous studies in this field.

This research is considered as a starting point to conduct in-depth analysis of CSFs in HEIs to increase the success rate of ERP implementation. Furthermore, it will enrich the academic knowledge in this field because of the lack of previous research on the successful implementation of ERP systems in the HEI sector. The researchers intend to carry on the reach by conducting in-depth analysis of different universities in Saudi Arabia that have implemented an ERP system.

REFRENCES

[1] Ifinedo, Princely (2011) "Examining the Influences of External Expertise and In-House Computer/IT Knowledge on ERP System Success",Journal of Systems and Software Vol. 84 No. 12, pp. 2065-2078

[2] Klaus, Helmut., Rosemann, Michael & Gable, Guy (2000) "What is ERP?", Information systems frontiers Vol. 2, No.2, pp. 141-162

[3] Mabert, VincentA., Sony, Ashok & Venkataramanan, Munirpallam (2003) "The Impact of Organization Size on Enterprise Resource Planning (ERP) Implementations in the US Manufacturing Sector", Omega 31(3), pp. 235-246

[4] Wang, Eric T.G., Shih, S.P.& Jiang, J.J. & Klein, G. (2008) "The Consistency among Facilitating Factors and ERP Implementation Success: A Holistic View of Fit". Journal of Systems and Software Vol. 81, No. 9, pp. 1609-1621

[5] Xu, LauraXiao Xia., Yu, Wang Feng, Lim, Roland & Hock, Lua Eng (2010) "A Methodology for Successful Implementation of ERP in Smaller Companies", In: 2010 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI).Qingdao, China, pp. 380-385.

[6] Siriginidi, Subba Rao (2000) "Enterprise Resource Planning in Reengineering Business",Business Process Management Journal,Vol. 6, No. 5, pp. 376-391.

[7] Al-Shamlan, Hala M. & Al-Mudimigh, Abdullah S. (2011)"The Change Management Strategies and Processes for Successful ERP Implementation: A Case Study of MADAR", International Journal of Computer Science,Vol. 8, No. 2, pp. 399-407.

[8] Liang, Huigang., Saraf, Nilesh., Hu, Qing. & Xue,Yajiong (2007) "Assimilation of Enterprise Systems: The Effect of Institutional Pressures and Mediating Role of Top Management", MISQuarterly, Vol. 31, NO. 1, pp. 59-87.

[9] Martin, M.H. (1998). "An ERP strategy". Time Inc., New York 137, PP. 95-97.

[10] Samuel, R. Dhinakaran & Kumar, Santhosh (2013) "Prediction of ERP Success Before the Implementation", In:International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012), Proceedings. Springer Berlin Heidelberg, pp. 219-227.

[11] Abugabah, Ahed & Sanzogni, Louis (2010) "Enterprise Resource Planning (ERP) System in Higher Education:A Literature Review and Implications", International Journal of Human and Social Sciences, Vol. 5, NO. 6, pp. 395-399.

[12] Seo, Goeun (2013) "Challenges in Implementing Enterprise Resource Planning (ERP) System in Large Organizations: Similarities and Differences Between Corporate and University Environment", Master’s thesis, Massachusetts Institute of Technology

[13] Al-Mashari, Majed., Al-Mudimigh, Abdullah. & Zairi, Mohamed (2003) "Enterprise Resource Planning: A Taxonomy of Critical Factors",European journal of operational research,Vol.146, No. 2, pp. 352-364.

[14] Ngai, Eric. WT., Law, Chuck CH& Wat & Francis KT (2008) "Examining the Critical Success Factors in the Adoption of Enterprise Resource Planning",Computers in Industry,Vol.59, No. 6, pp. 548-564.

[15] Zhang ,Zhe., Lee,Matthew K.O., Huang, Pei., Zhang, Lliang. &Huang, Xiaoyuan (2005) "A Framework of ERP Systems Implementation Success in China:An Empirical Study", International Journal of Production Economics, Vol. 98, No. 1, pp. 56-80.

[16] Bhatti, T. R. (2005) "Critical Success Factors for the Implementation of Enterprise Resource Planning (ERP): Empirical Validation", In:2nd International Conference on Innovation in Information Technology IIT press, Dubai, UAE, pp. 1-10.

[17] Holland Christopher P. & Light, Ben (1999) "ACritical Success Factors Model for ERP Implementation", IEEE software, Vol. 16, No. 3, pp. 30-36.

[18] Jafari, S.M., Osman, M.R., Yusuff, R.M. &Tang, S.H. (2006) "ERP Systems Implementation in Malaysia:Importance of Critical Success Factors", International Journal of Engineering and Technolog, Vol. 3, No.1, pp. 125-131.

[19] Kamhawi, Emad .M. (2007) "Critical Factors for Implementation Success of ERP Systems: An Empirical Investigation from Bahrain",International Journal of Enterprise Information Systems,Vol. 3, No. 2, pp. 34-49.

[20] Liu, Pang-Lo (2011) "Empirical Study on Influence of Critical Success Factors on ERP Knowledge Management on Management Performance in High-Tech Industries in Taiwan",Expert Systems with Applications, Vol.38, No. 8, pp.10696-10704.

[21] Loh, Tee Chiat& Koh, S.C.L. (2004) "Critical Elements for a Successful Enterprise Resource Planning Implementation in Small- and Medium-Sized Enterprises",International journal of production research, Vol. 42, No. 17, pp. 3433-3455.

[22] Motwani, Jaideep., Subramanian, Ram. & Gopalakrishna, Pradeep (2005) "Critical Factors for Successful ERP Implementation: Exploratory Findings from Four Case Studies", Computers in Industry , Vol. 56, No. 6, pp. 529-544.

[23] Soja, Piotr (2006) "SuccessFactors in ERP Systems Implementations: Lessons from Practice", Journal of enterprise information management, Vol. 19, No. 6, pp.646--661.

[24] Umble,Elisabeth J., Haft, Ronald. R. & Umble Michael M. (2003) "EnterpriseResource Planning: Implementation Procedures and Critical Success Factors", European journal of operational research , Vol. 146, No. 2, pp.241-257.

[25] Gates, Kathryn F. (2004) "Evaluating the North American Pilot for SAP's Campus Management System", In von Hellens, L., Nielsen, S., Beekhuyzen, J. (eds.), Qualitative Case Studies on Implementation of Enterprise Wide Systems. Idea Group, Hershey, pp. 192-210.

[26] Moon, Young B. (2007) "Enterprise Resource Planning (ERP): AReview of the Literature", International Journal of Management and Enterprise Development, Vol. 4, No. 3, pp. 235-264.

[27] Nielsen, Jens L. (2005) "Critical Success Factors for Implementing ERP System", In von Hellens, L., Nielsen, S., Beekhuyzen, J. (eds.), Qualitative case studies on implementation of enterprise wide systems. Idea Group, Hershey, pp. 211-231.

[28] Rabaa'i, Ahmad A. (2009) "Identifying Critical Success Factors of ERP Systems at the Higher Education Sector",In: ISIICT2009: Third International Symposium on Innovation in Information & Communicaton Technology, Philadelphia Uni., Amman, Jordan .

[29] Ahmad, Raja Lope, Othman, Zulkifli & Mukhtar, Mohsin (2011) "ERP Implementation Framework for Malaysian Private Institution of Higher Learning", In:2011 International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, pp. 1-5.

[30] Aldayel, AbeerI., Aldayel, Mashael S. & Al-Mudimigh, Abdullah S. (2011) "The Critical Success Factors of ERP Implementation in Higher Education in Saudi Arabia: A Case Study", journal of Information Technology & Economic Development , Vol. 2, No, 2, pp. 1-16.

[31] Ministry of Higher Education, (2015) [Online]. Available:
http://he.moe.gov.sa/ar/about/egovinitiaves/Pages/cInitiatives.aspx

[32] Gable, Guy G. (1998) "Large Package Software—A Neglected Technology?", Journal of Global Information Management, Vol.6, pp. 3-4.

[33] Rosemann, Michael & Wiese, Jens (1999) "Measuring the Performance of ERP Software—A Balanced Scorecard Approach", In:10th Australasian Conference on Information Systems, Wellington, New Zealand, pp. 733-784.

[34] Huang, Albert, Yen, David C., Chou, David C. & Xu, Yurong (2003) "Corporate Applications Integration: Challenges, Opportunities, and Implementation Strategies",Journal of Business and Management,Vol. 9 No. 2, pp. 137-145.

[35] Rashid, MohammadA., Hossain, Liaquat & Patrick, Jon David (2002) "The Evolution of ERP Systems: A Historical Perspective", In L. Hossain, J. Patrick, & M. Rashid, Enterprise Resource Planning: Global Opportunities & Challenges, Idea Group, United States of America, pp. 1--16.

[36] Beheshti, Hooshang. M. (2006) "What Managers Should Know about ERP/ERP II", Management Research News ,Vol. 29 No. 4, pp.184--193.

[37] Chang, Man-Kit., Cheung, Waiman., Cheng, C. Hung & Yeung, Jeff H.Y. (2008) "Understanding ERP System Adoption from the User's Perspective", International Journal of Production Economics,Vol. 113, No. 2, pp. 928--942.

[38] Gargeya, Vidyaranya B. &Brady, Cydnee (2005) "Success and Failure Factors of Adopting SAP in ERP System Implementation", Business Process Management Journal,Vol.11, No. 5, pp. 501-516.

[39] Luo, Wenhong & Strong, Diane. (2004) "A Framework for Evaluating ERP Implementation Choices", Engineering Management, IEEE Transactions,Vol. 51, No. 3, pp. 322-333.

[40] Spathis, Charalambos & Ananiadis, John (2005) "Assessing the Benefits of Using an Enterprise System in Accounting Information and Management", Journal of Enterprise Information Management, Vol. 18, No. 2, pp.195-210.

[41] Nah, Fui-Hoon Fiona, Lau, Janet Lee-Shang & Kuang, Jinghua (2001) "CriticalFactors for Successful Implementation of Enterprise Systems", Business process management journal, Vol. 7, No. 3, pp. 285-296.

[42] Poston, Robin & Grabski, Severin (2001) "Financial Impacts of Enterprise Resource Planning Implementations", International Journal of Accounting Information Systems,Vol. 2, No. 4, pp. 271-294 .

[43] Shang, Shari & Seddon, Peter B. (2000) "A Comprehensive Framework for Classifying the Benefits of ERP Systems", In: Proceedings of AMCIS 2000, Long Beach, CA, pp. 1005-1014.

[44] Shang, Shari & Seddon, Peter B. (2002) "Assessing and Managing the Benefits of Enterprise Systems: The Business Manager's Perspective", Information systems journal , Vol. 12, No. 4, pp. 271-299.

[45] Dezdar, Shahin & Sulaiman, Ainin (2009) "Successful Enterprise Resource Planning Implementation: Taxonomy of Critical Factors", Industrial Management & Data Systems, Vol. 109, No. 8, pp. 1037-1052.

[46] Xue, Yajiong, Liang, Huigang., Boulton, William R. & Snyder, Charles A. (2005) "ERP Implementation Failures in China: Case Studies with Implications for ERP Vendors",International journal of production economics, Vol. 97, No. 3, pp.279-295.

[47] Traci, Barker & Mark, N. Frolick (2003) "ERP Implementation Failure: A Case Study", Information Systems Management,Vol. 20, No. 4, pp. 43-49.

[48] Chen, Charlie C., Law, Chuck & Yang, Samuel C. (2009) "Managing ERP Implementation Failure: A Project Management Perspective",Engineering Management IEEE Transactions, Vol. 56, No. 1, pp. 157-170.

[49] Markus, M. Lynne, Axline, Sheryl., Petrie, David & Tanis, S. Cornelis. (2000) "Learning from Adopters' Experiences with ERP:Problems Encountered and Success Achieved", Journal of information technology , Vol. 15, No. 4, pp. 245-265.

[50] Pollock, Neil & Cornford, James (2004) "ERP Systems and the University as a "Unique"Organisation", Information technology & people, Vol. 17, No. 1, pp.31-52.

[51] Markus, M. Lynne & Tanis, Cornelis (2000) "The Enterprise Systems Experience—From Adoption to Success", Framing the domains of IT research: Glimpsing the future through the past 173, pp. 207-173.

[52] Rockart, John F. (1978) "Chief Executives Define Their Own Data Needs", Harvard business review, Vol. 57, No. 2, pp. 81-93.

[53] Rockart, John F. & Bullen, Christine V. (1986) "The Rise of Managerial Computing: The Best of the Center for Information Systems Research", Dow Jones-Irwin, New York.

[54] Pinto, Jeffrey K. & Slevin, Dennis P. (1987) "Critical Factors in Successful Project Implementation", Engineering Management, IEEE Transactions on 1 , Vol. 34, No. 1, pp. 22-27.

[55] Shanks, G., Parr, A.,Hu, B., Corbitt, B., Thanasankit, T.& Seddon, P. (2000) "Differences in Critical Success Factors in ERP Systems Implementation in Australia and China: A Cultural Analysis",In: Proceedings of ECIS 2000, Vienna, Austria, pp. 537–544.

[56] Somers, Toni M. & Nelson, Klara G. (2004) "A Taxonomy of Players and Activities Across the ERP Project Life Cycle", Information & Management, Vol. 41, No. 3, pp. 257-278.

[57] Finney, Sherry & Corbett, Martin (2007) "ERP Implementation: A Compilation and Analysis of Critical Success Factors", Business Process Management Journal, Vol. 3, No. 3, pp.329-347.

[58] Thomas, Glenn A. & Jajodia, Shyam (2004) "Commercial-Off-the-Shelf Enterprise Resource Planning Software Implementations in the Public Sector: Practical Approaches for Improving Project Success", The Journal of Government Financial Management , Vol. 53, No. 2, pp.12--19.

[59] Zornada, Leo & Velkavrh, TamaraBertok (2005) "Implementing ERP Systems in Higher Education Institutions",In: 27th International Conference on Information Technology Interfaces ICTI, Cavtat, Croatia.

[60] Lockwood, G. (1985) "Universities as Organizations", In Lockwood, G. and Davies, J. (eds), Universities; The Management Challenge, Windsor, UK., pp. 139-163.

[61] Heiskanen, Ari., Newman, Michael& Similä, Jouni (2000) "The Social Dynamics of Software Development", Accounting, Management and Information Technologies, Vol. 10, No. 1, pp. 1-32.

[62] Bologa, Razvan., Bologa, Ana-Ramona & Sabau, Gheorghe (2009) "Success Factors for Higher Education ERPs",International Conference on Computer Technology and Development, pp. 28-32.

[63] King, Paula., Kvavik, Robert B. & Voloudakis, J. (2002) "Enterprise Resource Planning Systems in Higher Education",EDUCAUSE Center for Applied Research: Research Bulletin 22, pp. 1-11.

[64] Sabau, G., Munten, M., Bologa, A. R., Bologa, R. & Surcel, T. (2009) "An Evaluation Framework for Higher Education ERP Systems", WSEAS Transactions on Computers, Vol. 8, No. 11, pp. 1790-1799.

[65] Beekhuyzen, J., Goodwin, M., Nielsen, J.L. & Uervirojnangkoorn, M. (2001) "ERP Implementation at Australian Universities", Technical Report, Brisbane, Australia, Griffith University.

[66] von Hellens, Liisa, Nielsen, Sue & Beekhuyzen, Jenine (2005) "Qualitative Case Studies on Implementation of Enterprise Wide Systems", IGI Global, Hershey.

[67] Davison, Robert. (2002) "Cultural Complications of ERP", Communications of the ACM,Vol. 45, No.7, pp.109-111.

[68] Parth, Frank R., Gumz, J. (2003) "Getting Your ERP Implementation Back on Track".

[69] Addo-Tenkorang, Richard & Helo, Petri (2011) "Enterprise Resource Planning (ERP): A Review Literature Report",Proceedings of the World Congress on Engineering and Computer Science, San Francisco, CA, Vol IIpp19-21.

[70] Asemi, Asefeh & Jazi, Mohammad (2010) "A Comparative Study of Critical Success Factors (CSFs) in Implementation of ERP in Developed and Developing Countries",International Journal, Vol. 2, No.5, pp99-110.

[71] Esteves, Jose & Bohórquez, Victor (2007) "An updated ERP systems annotated bibliography: 2001–2005", Instituto de Empresa Business School. Working Paper No. WP, Vol. 7, No. 4M pp 2-59.

[72] Shehab, Essam, Sharp,Supramaniam & Spedding (2004) "Enterprise resource planning: An integrative review",Business Process Management Journal, Vol. 10, No. 4, pp359-386.

[73] Botta-Genoulaz, Valerie, Millet, Pierre-Alain & Grabot, Bernard (2005) "A survey on the Recent Research Literature on ERP Systems", Computers in Industry, Vol. 56, No. 6, pp510-522.

[74] Rabaa'i, Ahmad A., Bandara, Wasana & Gable, Guy (2009) "ERP Systems in the Higher Education Sector: ADescriptive Study",In:Proceedings of the 20th Australasian Conference on Information Systems, Monash Uni.: Caulfield Campus, Melbourne, Australia, pp. 456-470.

[75] Allen, David., Kern, Thomas & Havenhand, Mark (2002) "ERP Critical Success Factors: An Exploration of the Contextual Factors in Public Sector Institutions", In:Proceedings of the 35th Annual Hawaii International Conference on System Sciences(HICSS), Hawaii,U.S.A, pp. 3062-307.

[76] Holland, Christoper P., Light, Ben & Gibson, Nicola (1999) "A Critical Success Factors Model for Enterprise Resource Planning Implementation", Proceedings of the 7th European Conference on Information Systems, Vol. 1, pp. 273-287.

[77] Chatfield, Craig (2005) "Factors that Affect ERP System Success",Qualitative Case Studies on Implementation of Enterprise Wide Systems,L. von Hellens,S. Nielsen, & J. Beekhuyzen,Eds.Hershey, Pennsylvania: IGI, pp232-242.

[78] Olugbara, O.O., Kalema, B.M. & Kekwaletswe, R.M. (2014) "Identifying Critical Success Factors: The Case of ERP Systems in Higher Education", The African Journal of Information Systems, Vol. 6, No. 3, pp. 65-84.

[79] Somers, Toni M. & Nelson, Klara (2001) "The Impact of Critical Success Factors Across the Stages of Enterprise Resource Planning Implementations",In Proceedings of the 34th Annual Hawaii International Conference on System Sciences, Maui, Hawaii, pp. 1-10.

[80] Nah, F.F.H. & Delgado, Santiago (2006) "Critical Success Factors for Enterprise Resource Planning Implementation and Upgrade", Journal of Computer Information Systems , Vol. 46, No. 5, pp. 99-113.

[81] E-Government Program (2015) [Online]. Available: http://www.yesser.gov.sa/en/ProgramDefinition/Pages/Overview.aspx

[82] Al-Hudhaif, Sulaiman A. (2012) "ERP Implementation at King Saud University". Global Journal of Management and Business Research,Vol. 12, No. 5, pp. 71-77.

[83] Alghathbar, Khaled (2008) "Practical ERP Implementation in Government Organization",International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, Las Vegas, USA, pp. 343-349.

[84] Frye, Doug., Gulledge, T., Leary, M. & Sommer, R. (2007) "PublicSector Enterprise System Implementation", Electronic Government, An International Journal, Vol. 4, No. 1, pp. 76-96.

[85] Rabaa'i, Ahmad A. (2009) "The Impact of Organisational Culture on ERP Systems Implementation: Lessons from Jordan", In:Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2009), Hyderabad, India.

AUTHORS

[1]Ashwaq Sulaiman AlQashami earned her bachelor degree with a specialization in Information Systems at the College of Computer and Information Sciences at Al-Imam Muhammad Ibn Saud University in Riyadh, Saudi Arabia, in 2009. She is currently pursuing a master's degree from the same department with an expected graduation date in 2016. She has Six years of experience in both public and private sectors in systems analysis, quality assurance and training. Currently, she is working as a programmer at Princess Nora University. Her current research interests include enterprise information systems, information technology governance and project management.

[2]Heba Mohammad is an Assistant Professor of Information Systems at the College of Computer and Information Sciences of Al-Imam Muhammad Ibn Saud University. She received her Ph.D in e-business from University of Salento, Italy. Her research focuses on enterprise systems, knowledge management, communities of practice, e-business and e-learning. She also provides different consultation services to various institutions.

MOCANAR: A MULTI-OBJECTIVE CUCKOO SEARCH ALGORITHM FOR NUMERIC ASSOCIATION RULE DISCOVERY

Irene Kahvazadeh and Mohammad Saniee Abadeh

Faculty of Electrical and Computer Engineering,
Tarbiat Modares University, Tehran, Iran
i.kahvazadeh@modares.ac.ir
saniee@modares.ac.ir

ABSTRACT

Extracting association rules from numeric features involves searching a very large search space. To deal with this problem, in this paper a meta-heuristic algorithm is used that we have called MOCANAR. The MOCANAR is a Pareto based multi-objective cuckoo search algorithm which extracts high quality association rules from numeric datasets. The support, confidence, interestingness and comprehensibility are the objectives that have been considered in the MOCANAR. The MOCANAR extracts rules incrementally, in which, in each run of the algorithm, a small number of high quality rules are made. In this paper, a comprehensive taxonomy of meta-heuristic algorithm have been presented. Using this taxonomy, we have decided to use a Cuckoo Search algorithm because this algorithm is one of the most matured algorithms and also, it is simple to use and easy to comprehend. In addition, until now, to our knowledge this method has not been used as a multi-objective algorithm and has not been used in the association rule mining area. To demonstrate the merit and associated benefits of the proposed methodology, the methodology has been applied to a number of datasets and high quality results in terms of the objectives were extracted.

KEYWORDS

Numeric Association Rule, Cuckoo Search, Multi-Objective Algorithm

1. INTRODUCTION

Association rule mining methods are one of the most used methods to extract relationships among features of a dataset; They were introduced in [1]. An association rule, denoted by X→Y is defined with two parts, antecedent part (X) and consequent part (Y) and both of them contain an item set. There are many impressive methods to obtain association rules in various applications [2-4], although most of them require values of the features to be discrete. For this reason these techniques discretize the numeric features but this causes a loss of information. If we want to discover association rules from continuous features, we should deal with a large search space since when the features are continuous, the number of the association rules can be

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 99–113, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51509

discovered are numerous. To solve the problem of large search space, one of the best suggestions is to use meta-heuristic algorithms.

The meta-heuristics are divided into two categories according to our knowledge, biological and Bio_Inspired Algorithms that are illustrated in Figure 1. The meta-heuristic methods are usually based on a physical phenomenon or based on the biological methods such as Simulated Annealing as suggested in [5], Gravitational Search Algorithm [6], Magnetic Optimization Algorithm [7], External Optimization [8] and Harmony Search [9]. We divide the Bio_Inspired meta-heuristics into two categories, evolutionary methods that use Darwin's theory directly, such as Genetic Algorithm that is suggested in [10], Genetic Programming [11], Evolution Strategy [12], Evolutionary Programming [13] and so on. Swarm intelligence based methods are other Bio_Inspired meta-heuristic algorithms that are mainly inspired by lives of living organisms. Swarm intelligence based methods also can be divided into Stigmergic based and imitation based categories. The Stigmergic based methods use an environmental memory to establish communication indirectly. The pheromone table in Ant Colony Optimization (ACO) is an example of environmental memory. ACO is suggested in [14], Honeybee Hive Optimization [15] and Termite Colony Optimization [16] are examples of the Stigmergic based methods. The imitation based methods have not any shared environmental memory and the communication between them is directly. All individuals in imitation based methods have a local memory and a global memory. The individuals are desired to the local best and global best positions. Particle Swarm Optimization [17], Imperialist Competitive Algorithm [18], Firefly Algorithm [19], Shuffled Frog-Leaping [20], Cat Swarm Optimization [21], Fruit Fly Optimization [22], Bacterial Foraging Optimization [23], Artificial Fish Swarm Algorithm [24], Bat Algorithm [25], Lion Pride Optimizer [26], Krill Herd Algorithm [27], Hunting Search [28] and Cuckoo Optimization Algorithm [29] are some examples of these methods. Mentioned taxonomy [30], The taxonomy is demonstrated in Figure 1.

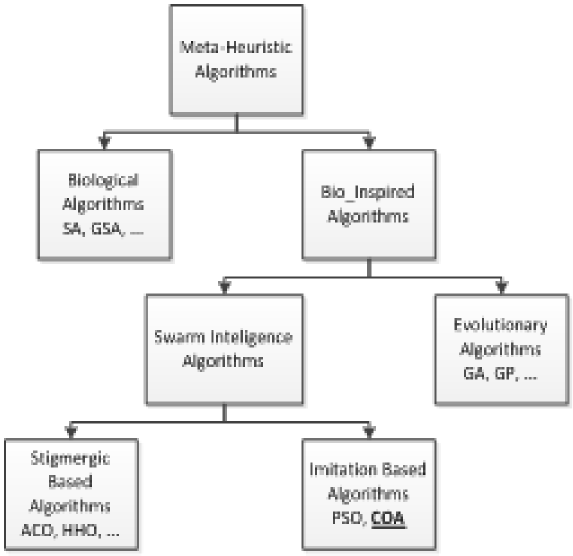

Meta-Heuristic Algorithms
Biological Algorithms SA, GSA, ...
Bio_Inspired Algorithms
Swarm Inteligence Algorithms
Evolutionary Algorithms GA, GP, ...
Stigmergic Based Algorithms ACO, HHO, ...
Imitation Based Algorithms PSO, COA

Figure 1. Meta-Heuristic Algorithms Categories [30]

If we use older algorithms, they may not have the capabilities of new algorithms; but however, they have been used in various applications and their performance is guaranteed. Choosing very new algorithms can be tricky, because they have not been used extensively and they might have unknown drawbacks. Many meta-heuristic algorithms have been used to discover association rules like [31-36] but in extracting of the numeric association rules, we have use Cuckoo Search [29] because this algorithm is one of the most matured algorithms and also, it is a simple and understandable algorithm. In addition, until now this method has not been used as a multi-objective algorithm and has not been used in the association rule mining area.

The paper is organized as follows: In the next section we describe preliminaries and in Section 3, the proposed method is explained. Section 4 contains experimental results and discussion and the last section concludes the paper.

2. PRELIMINARIES

This section consists of two subsections. In the first one, multi-objectivity concepts are described. Multi-objective approaches are divided into three categories and are explained separately. Also our objectives that are considered for numeric association rule mining are described in this subsection. In the second subsection, we discuss about the cuckoos life and review the studies that are inspired from their life.

2.1. Multi-Objectivity

Usually, the multi-objective problems are solved with one of the three multi-objective approaches: aggregation based approach, population based approach and the Pareto based approach. These approaches have their advantages and disadvantages. We studied these approaches in the following.

Aggregation based approach: in this approach, all of the objectives are combined into one objective which is done using mathematical operators like subtracting, multiplying and so on. An example is offered in (1) for this approach that uses the sum operator and weights for objectives.

f(x)= w_1*f_1 (x)+w_2*f_2 (x)+⋯+w_k*f_k (x)

where x∈ X_f (1)

w_iis the weight of ith objective, k is the number of objectives and X_fis the search space. This approach is one of the easiest approaches but the weights must be properly defined. One of the disadvantages of this approach is that the approach cannot discover the concave parts of the Pareto front [37]. Nonlinear aggregation functions do not have this restriction. One of the studies that used this approach is [38].

Population based approach: in this approach, the population is divided into k (that is the number of objectives) sub-populations. Each sub-population is improved with regards to one objective and finally after termination of the algorithm the sub-populations are aggregated in one solution to the k-objective problem. Since this approach is easy to use, it is well-known among the researchers, [39] is one of the studies that use a population based approach to solve multi-objective problems.

Pareto based approach: It rarely happens to have a unique solution that is optimal in terms of all objectives. So instead of looking for a unique solution that is optimal, we should trade-off between the objectives. Pareto optimality definition says that a solution is a Pareto optimal, if there exists no feasible solution in X_f which would improve some objective without causing a simultaneous deterioration in at least one other objective. [40-43] studies have used Pareto based approach. Also in this study we use a Pareto based approach that considers four objectives: support, confidence, interestingness and comprehensibility. These objectives are important in association rule mining area. The objectives are defined as follows: The support of an item set X, denoted by S(X), is the ratio of the number of records ($|R_X|$) that contains the item set X to the total number of records ($|D|$). S(X) is defined by (2). The support of an association rule is denoted by $S(X \rightarrow Y)$ and is the ratio of the number of records containing both X and Y ($|R_X \cup R_Y|$), to the total number of records, $|D|$. If the support of an association rule is 20%, this means that 20% of the analyzed records contain $X \cup Y$. $S(X \rightarrow Y)$ is defined by (3).

$$S(X)=|R_X|/|D| \qquad (2)$$

$$S(X \rightarrow Y)=|R_X \cup R_Y|/|D| \qquad (3)$$

The confidence of an association rule indicates the degree of correlation between X and Y in the dataset. The confidence of an association rule denoted by $C(X \rightarrow Y)$ is the ratio of the number of records that contain $X \cup Y$ to the number of records that contain X. If we say an association rule has a confidence of 80%, it means that 80% of the records containing X also contain Y. The confidence of an association rule is defined by (4).

$$C(X \rightarrow Y)=S(X \rightarrow Y)/S(X) = |R_X \cup R_Y|/|R_X| \qquad (4)$$

In addition to support and confidence measures, two other measures are used to mine high quality association rules. If the number of conditions involved in the antecedent part is less than the number of conditions in the consequent part, the rule is more comprehensible [44]. The comprehensibility is computed by (5).

$$Comp.= \log(1+|R_X|)/\log(1+|R_X \cup R_Y|) \qquad (5)$$

Interestingness measure refers to finding rules that are interesting or useful, not just all possible rules. In some approaches, to find interestingness the entire dataset is divided based on each feature presented in the consequent part. Since different numbers of features can present in the consequent part and because they are not predefined, this approach may not be feasible for association rule mining. So, a new expression is defined in [45] which uses the support count of the antecedent and the consequent parts of the rules. This expression is shown in (6).

$$Inter.=|R_X \cup R_Y|/|R_X| * |R_X \cup R_Y|/|R_Y| * (1-|R_X \cup R_Y|/|D|) \qquad (6)$$

The equation contains three parts. The first expression describes probability of generating the rule based on the antecedent part. The second expression shows the probability based on the consequent part, and the last one $(1-|R_X \cup R_Y|/|D|)$ describes the probability of not generating the rule based on the whole dataset.

2.2. Cuckoo's Life

Some of birds are known as Brood Parasites. These birds instead of having to build their own nest, lay eggs in the nests of other birds and so the owner of the nest takes care of the Brood Parasites eggs. The cuckoo is one of the most famous Brood Parasite birds and is an expert in deception of the other birds. The female cuckoo destroys one of the other bird's eggs and replaces her egg. Every bird differs in the color and pattern of its egg but the cuckoos have an uncanny talent for mimicry. This talent is one of the mysteries of nature. Of course, some of host birds know the stranger egg and they destroy it or they leave their nest forever. In fact, the cuckoos boosting their mimic power and the host birds boosting their identification power and these effort and fight are an incessant matter. Various types of algorithms in various applications have been proposed which are inspired by the cuckoo's life like [46-50] but our method is very similar to [29]. Details of our method will be explained in Section 3. In the paper, we suggest the multi-objective version of cuckoo search in the numeric association rule mining context for first time.

3. PROPOSED METHOD

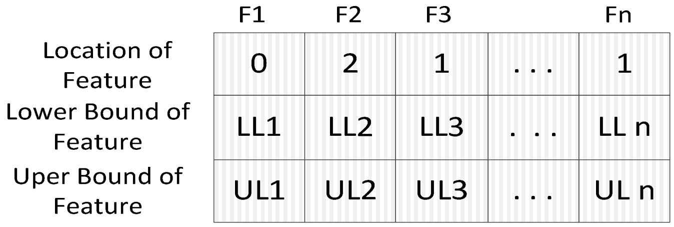
In this section our proposed method is explained. This section consists of two subsections. In the first one, representation of the numeric association rules with cuckoo search algorithm is demonstrated and in the next one, MOCANAR is explained in detail.

3.1. Representation of Problem

In this paper, the cuckoos are represented with 2D array for association rule mining problem that is illustrated in Figure 2. The number of columns of the array is equal to n that shows the number of features in dataset and the number of rows is equal to 3 that first one shows location of each feature in current association rule, the second row shows the lower bound of the feature value and the third row shows the upper bound of the feature value in the current association rule. If the value of a cell in first row is set to 0, the related feature is not present in the association rule and if a cell in first row contains 1, it means that the related feature is in the antecedent part of the association rule and the value 2 in the cell shows the related feature is in the consequent part of the current association rule. For example, the Figure 2 shows the following association rule:

if ($LL3<F3<UL3$ and $LLn<Fn<ULn$)

then ($LL2<F2<UL2$)



	F1	F2	F3		Fn
Location of Feature	0	2	1	. . .	1
Lower Bound of Feature	LL1	LL2	LL3	. . .	LL n
Uper Bound of Feature	UL1	UL2	UL3	. . .	UL n

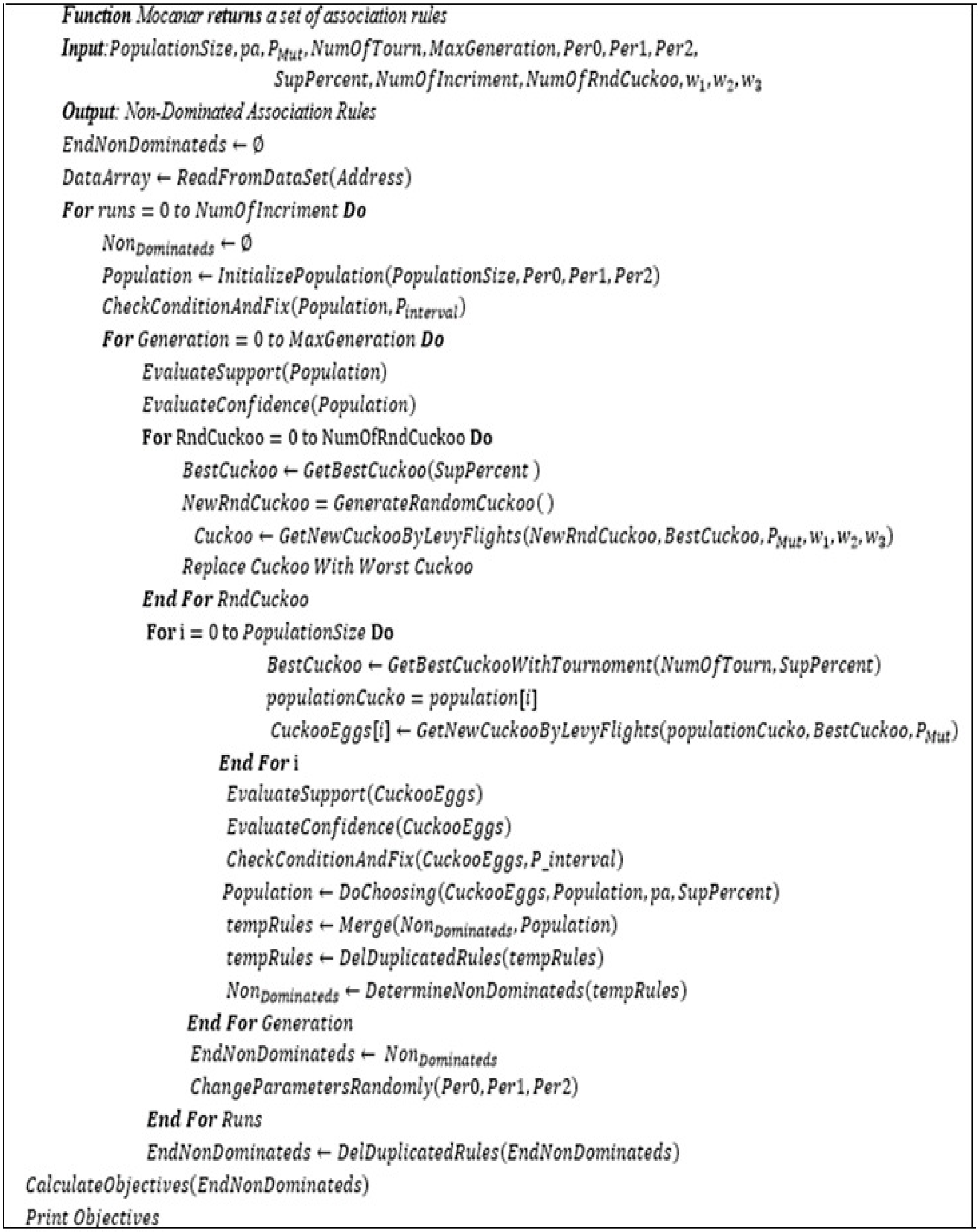
Figure 2. Representation of the Association Rules with Cuckoos.

3.2. MOCANAR: Multi-objective Cuckoo Search for Numeric Association Rule Mining

MOCANAR is a multi-objective cuckoo search algorithm that extracts high quality association rules from numeric datasets. The support, confidence, interesting and comprehensibility are the objectives that are considered in the MOCANAR. The MOCANAR extracts rules incrementally in which, in the each increment, low numbers of high quality rules are made. The number of increments is determined by NumOfIncrement parameter. To generate the low number of the high quality rules in each increment, an iterative loop is repeated NumGeneration (which is another input parameters) times. During the execution of these iterations that are called generations, the initial random association rules are improved in evolutionary way. Our chosen meta-heuristic method is the cuckoo search that intelligently improves the rules in generations. Each generation consists of two 'for-do' cycles. In the first one, since the convergence of the algorithm is very fast, NumOfRndCuckoo (another input parameter) numbers of random cuckoos are generated and are directed toward the best cuckoo by using the levy flight policy and so are replaced with worst cuckoos in the population. In the second 'for-do' cycle, each cuckoo in the population generates an egg by using the levy flights and so pa percent of the generated eggs are eliminated. Pseudo code of the MOCANAR is illustrated in Figure 3 and is explained in detail below. In the pseudo code, FinalNonDominateds keeps the non-dominated association rules from last increment. When the increments are finished, the FinalNonDominateds contains the final non-dominated association rules which will be shown to the user. The non-dominated association rules that are achieved in generations are stored in Non_Dominateds. DataArray stores the dataset. The increments are started from line 6. In line 8, the population is initialized by the InitializePopulation function. In this function, PopulationSize is the parameters that specifies the number of cuckoos of the population, Per0 specifies the possibility of placing a value of zero, Per1 specifies the possibility of placing a value of 1 and Per2 specifies the possibility of placing a value of 2 in the first row of association rules for each features that illustrated in Figure 2. In line 9, CheckConditionAndFixfunction checks two defined conditions for association rules: the first one, there should be at least one feature in the antecedent part of the rule and at least one feature in the consequent part of the rule; the second condition says that the range of lower bound and upper bound of the normalized features should not be greater that F_interval(that is another input parameter). Because the rules should not be too general, the F_interval parameter is used. The statements that are in the 'for-do' statement in line 10, are executed for NumGeneration (that is one of the input parameters) times. EvaluateObjectives function in the pseudo code calculates our objectives. 'for-do' statement in line 12 is executed NumOfRndCuckoo (that shows number of random cuckoos in each generation) times. High value for NumOfRndCuckoo increases the exploration ability of the algorithm and low value increases the exploitation ability of the algorithm. GetBestCuckooWithTournament function, determines the best cuckoo with Pareto policy in terms of our objectives, in which NumOfTourn numbers. of the cuckoos in population are selected randomly and so non-dominated cuckoos in terms of our objectives are removed. One of the non-dominated rules is returned by GetBestCuckooWithTournament randomly. In line 15, the generated random cuckoos are directed to the selected best rule by levy flight policy. The Levy flight essentially provides a random walk while the random step length is specified by a Levy distribution that is shown in (1)

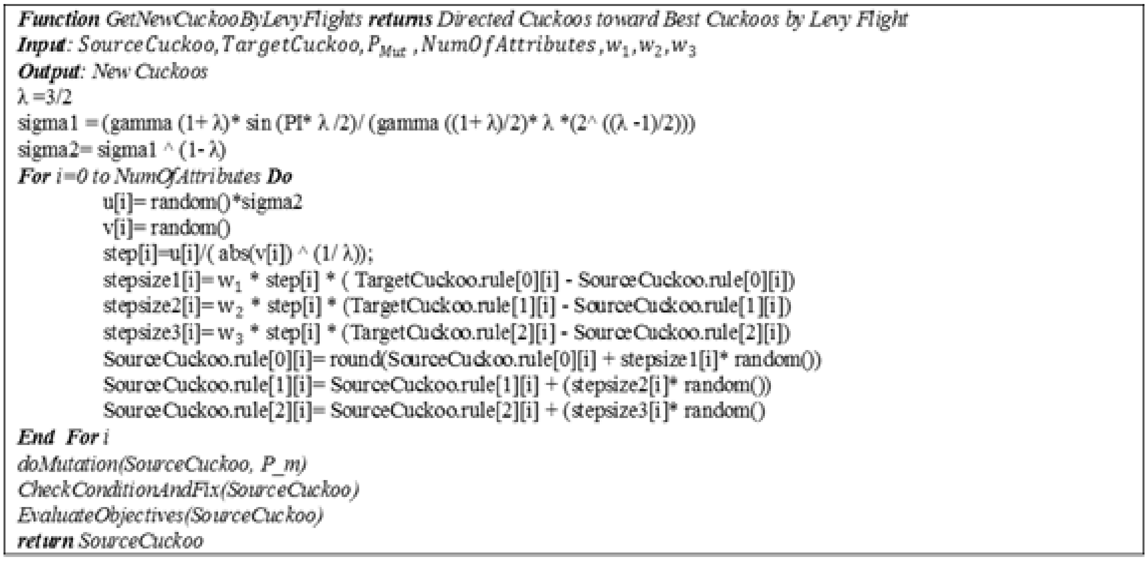
$$\text{Levy} \sim u=t^{(-\lambda)} \quad ,1<\lambda<3 \qquad (1)$$

The levy distribution has an infinite variance with an infinite mean. The implementation of this distribution to directing the cuckoos toward best cuckoo in detail is shown in Figure 4 for the readers that want to implement it. This directed rule is replaced with worst rule in the population. In lines 18-21, each of the cuckoos in population are directed to best cuckoo by using levy flight policy; the best cuckoo is selected with GetBestCuckooWithTournament function. The directed cuckoos are known as cuckoo eggs. These eggs should be checked in terms of the aforementioned conditions and should be evaluated in terms of our objectivities. DoChoosing function chooses PopulationSize numbers of the cuckoos in population and eggs in terms of our objectives and put them in the new population. This function is explained later with pseudo code. Current population is merged with last non-dominated rules by using Mergefunction and so the duplicated rules are deleted from them and later, the non-dominated rules are selected from them by using Pareto policy. This non-dominated rule set is related to generations and is different with the increment's non-dominated rule set. The generation's non-dominated rules are accumulated in increment's non-dominated rule set at the end of each increment. Then the Per0,Per1 and Per2 parameters are changed randomly to investigation of the other spaces of the search space in each increment. Changing those parameters helps to have different rules in each increment. Finally, the duplicated rules in FinalNonDominateds rule set are eliminated. In the Figure 4, SourceCuckoo is the cuckoo that should be directed toward TargetCuckoo (best cuckoo). NumOfAttributes parameter shows the number of dimensions of the cuckoo (the number of data features) and P_(Mut)parameter specifies the probability of mutation on each dimension of the cuckoo. w_1,w_2 and w_3 are the step sizes of cuckoo rule in each row of Figure 2 respectively which should be related to the scales of the problem of interests. In most cases, we can use 1 value for them. SourceCuckoo.rule in pseudo code refers to the 2D array shown in Figure 2. The rows of the array in each dimension are directed toward the TargetCuckoo. After this process, mutation operation is applied to each dimension by probability of P_(Mut). the resulted cuckoos are known as new eggs that should be checked in terms of aforementioned two conditions In line 24 of Figure 3, we have two populations (CuckooEggs,Population) and both of them have PopulationSize numbers of eggs and cuckoos respectively, and The DoChoosing function tries to select the PopulationSize numbers of them for the new population, since the size of population in the algorithm is constant. The Obj-share in the DoChoosing pseudo code shows the share of each objectives in new population. If its value is equal to 25 it means that 25 places in population are reserved per each objective. First, the eggs population is sorted with respect to support and so Pa percent of eggs in that population are deleted according to cuckoo search policies. Then two populations are merged in a population called tempRules; from here, the eggs are known as cuckoos. The tempRules are sorted with respect to support and so the Obj-share number of cuckoos are elected to be placed in the new population. The elected cuckoos are eliminated from tempRules.Also, this process is applied to confidence, interestingness and comprehensibility objectives. Finally, the new population is returned to main method. Here it can be said that we use the concepts of the Population based multi-objective approach.



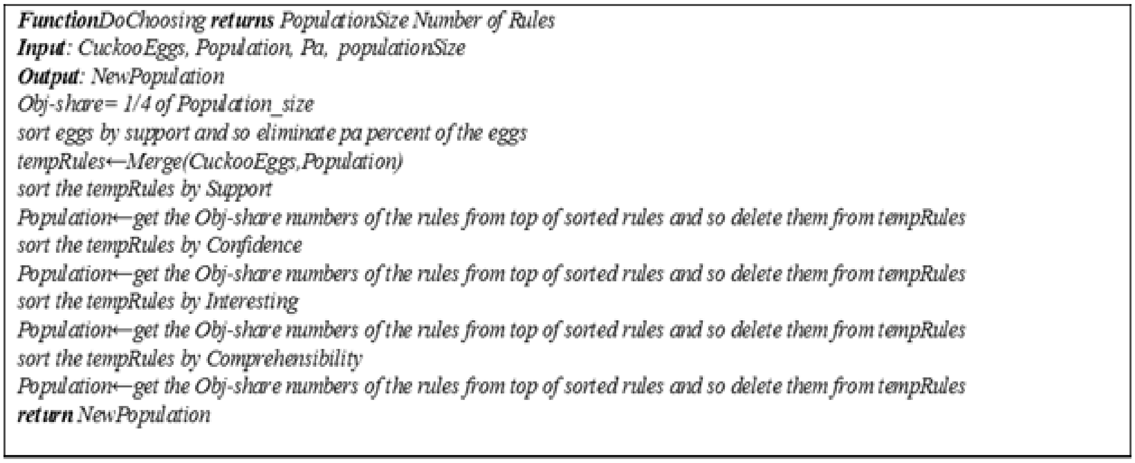
```
Function Mocanar returns a set of association rules
Input: PopulationSize, pa, $P_{Mut}$, NumOfTourn, MaxGeneration, Per0, Per1, Per2,
       SupPercent, NumOfIncriment, NumOfRndCuckoo, $w_1$, $w_2$, $w_3$
Output: Non-Dominated Association Rules
EndNonDominateds ← Ø
DataArray ← ReadFromDataSet(Address)
For runs = 0 to NumOfIncriment Do
    $Non_{Dominateds}$ ← Ø
    Population ← InitializePopulation(PopulationSize, Per0, Per1, Per2)
    CheckConditionAndFix(Population, $P_{interval}$)
    For Generation = 0 to MaxGeneration Do
        EvaluateSupport(Population)
        EvaluateConfidence(Population)
        For RndCuckoo = 0 to NumOfRndCuckoo Do
            BestCuckoo ← GetBestCuckoo(SupPercent )
            NewRndCuckoo = GenerateRandomCuckoo( )
            Cuckoo ← GetNewCuckooByLevyFlights(NewRndCuckoo, BestCuckoo, $P_{Mut}$, $w_1$, $w_2$, $w_3$)
            Replace Cuckoo With Worst Cuckoo
        End For RndCuckoo
        For i = 0 to PopulationSize Do
            BestCuckoo ← GetBestCuckooWithTournoment(NumOfTourn, SupPercent)
            populationCucko = population[i]
            CuckooEggs[i] ← GetNewCuckooByLevyFlights(populationCucko, BestCuckoo, $P_{Mut}$)
        End For i
        EvaluateSupport(CuckooEggs)
        EvaluateConfidence(CuckooEggs)
        CheckConditionAndFix(CuckooEggs, P_interval)
        Population ← DoChoosing(CuckooEggs, Population, pa, SupPercent)
        tempRules ← Merge($Non_{Dominateds}$, Population)
        tempRules ← DelDuplicatedRules(tempRules)
        $Non_{Dominateds}$ ← DetermineNonDominateds(tempRules)
    End For Generation
    EndNonDominateds ← $Non_{Dominateds}$
    ChangeParametersRandomly(Per0, Per1, Per2)
End For Runs
EndNonDominateds ← DelDuplicatedRules(EndNonDominateds)
CalculateObjectives(EndNonDominateds)
Print Objectives
```

Figure 3. Pseudo Code of the MOCANAR



```
Function GetNewCuckooByLevyFlights returns Directed Cuckoos toward Best Cuckoos by Levy Flight
Input: $SourceCuckoo, TargetCuckoo, P_{Mut}, NumOfAttributes, w_1, w_2, w_3$
Output: New Cuckoos
λ =3/2
sigma1 =(gamma (1+ λ)* sin (PI* λ /2)/ (gamma ((1+ λ)/2)* λ *(2^ ((λ -1)/2)))
sigma2= sigma1 ^ (1- λ)
For i=0 to NumOfAttributes Do
            u[i]= random()*sigma2
            v[i]= random()
            step[i]=u[i]/( abs(v[i]) ^ (1/ λ));
            stepsize1[i]= $w_1$ * step[i] * ( TargetCuckoo.rule[0][i] - SourceCuckoo.rule[0][i])
            stepsize2[i]= $w_2$ * step[i] * (TargetCuckoo.rule[1][i] - SourceCuckoo.rule[1][i])
            stepsize3[i]= $w_3$ * step[i] * (TargetCuckoo.rule[2][i] - SourceCuckoo.rule[2][i])
            SourceCuckoo.rule[0][i]= round(SourceCuckoo.rule[0][i] + stepsize1[i]* random())
            SourceCuckoo.rule[1][i]= SourceCuckoo.rule[1][i] + (stepsize2[i]* random())
            SourceCuckoo.rule[2][i]= SourceCuckoo.rule[2][i] + (stepsize3[i]* random()
End  For i
doMutation(SourceCuckoo, P_m)
CheckConditionAndFix(SourceCuckoo)
EvaluateObjectives(SourceCuckoo)
return SourceCuckoo
```

Figure 4: Directing the Cuckoo toward Best Cuckoo by Levy Function



```
FunctionDoChoosing returns PopulationSize Number of Rules
Input: CuckooEggs, Population, Pa,  populationSize
Output: NewPopulation
Obj-share= 1/4 of Population_size
sort eggs by support and so eliminate pa percent of the eggs
tempRules←Merge(CuckooEggs,Population)
sort the tempRules by Support
Population←get the Obj-share numbers of the rules from top of sorted rules and so delete them from tempRules
sort the tempRules by Confidence
Population←get the Obj-share numbers of the rules from top of sorted rules and so delete them from tempRules
sort the tempRules by Interesting
Population←get the Obj-share numbers of the rules from top of sorted rules and so delete them from tempRules
sort the tempRules by Comprehensibility
Population←get the Obj-share numbers of the rules from top of sorted rules and so delete them from tempRules
return NewPopulation
```

Figure 5: Pseudo Code of the Dochoosing Function

4. EXPERIMENTAL RESULTS AND DISCUSSION

We assess our proposed method in three public domain datasets: Basketball, Body fat and Quake. These datasets are available from the Bilkent University Function Approximation Repository[51]. Characteristics of the datasets are shown in Table 1 in which, second column shows the number of records in each dataset and third column shows the number of features for each dataset.

Table 1. Datasets Characteristics

Dataset	Number of Records	Number of Features
Basketball	96	5
Body fat	225	18
Quake	2178	4

All of the parameters of MOCANAR are described in the Section3. The used parameters values for each dataset are shown in Table 2. First row in Table 2 shows the name of datasets, the second row shows the size of MOCANAR population for each dataset. A high value for this parameter causes the algorithm to explore more of the search space but on the other side, it is time consuming. The third row shows the number of generations in each increment of algorithm. A high value for this parameter in addition to cause further explore in the search space, leads the algorithm to a better convergence. The pa parameter shows percentage of the cuckoo eggs that are scheduled to be eliminated in each generation. A low value for pa causes the algorithm less attention to previous generation cuckoos. The NumOfRndCuckoo parameter shows the number of the random cuckoos in each generation. A high value for this parameter increases the search space of the algorithm in comparing to its exploitation ability. The NumOfIncrement specifies the number of increments in the algorithm. A high value for this parameter increases the number of final non-dominated association rules. The NumOfTourn parameter shows the number of cuckoos that should be selected in tournament selection when the algorithm finds the best cuckoo. The P_Mut parameter determines the probability of the mutation after the eggs are generated. Increasing the value of this parameter causes increasing in the exploration ability of the algorithm and also, causes the algorithm to avoid local optimums.

The F_interval parameter specifies the maximum ranges between the Lower Limit (LL) and Upper Limit (UL) of the features. A high value for this parameter causes the generated association rules to be more general. In this study its value is set to 0.5 * (max value of feature - min value of feature). w_1,w_2 and 〖 w〗_3 that are used in GetNewCuckooByLevyFlights function, specify the length of steps in three rows of cuckoo (illustrated in Figure 2) to move toward the better position. High values for these parameters causes increasing in steps length and this larger steps leads to a faster convergence. In this paper, w_1,w_2 andw_3 values are equal to 1. Per0,Per1 and Per2 parameters are initialized randomly in which sum of thier values is equal to 1. The values of these parameters are changed in each increment randomly to produce different association rules in the increments. The proposed method is run 10 times and the results are averaged. In the following the results are shown and compared with other studies.

Table 2. Parameters Values for each Dataset

Dataset	Basketball	Quake	Body fat
$PopulationSize$	300	300	500
$NumGeneration$	300	300	250
pa	0.3	0.2	0.3
$NumOfRndCuckoo$	1	2	1
$NumOfIncrement$	40	50	50
$NumOfTourn$	30	50	100
P_{Mut}	0.05	0.2	0.1

In Tables 3, 4, 5, 7 and 8, the results obtained from our method are compared with results from Alatas and Akin[52],Alatas and Akin [53]and Minaei, Barmaki, and Nasiri[45]. Comparison in terms of the extracted rules count is shown in the Table 3. Increasing in the number of the extracted rules causes the increasing in the discovered knowledge but on the other side it decreases the interpretability of results. In Table 4, comparisons in terms of confidence are shown. The results show that in most cases, MOCANAR yields better results. Table 5 shows that MOCANAR has got best results in terms of the support measure. Because the algorithm runs many increments and in each increment tries to generate low number of high quality rules, MOCANAR has better support and confidence compared to other studies. Unfortunately, these increments are a little time consuming. The spent times for each dataset in 15 runs are shown in Table 6.

Table 3. Comparison in terms of Number of Association Rules

Dataset	Alatas [52]	RPSO [53]	MOGAR [45]	MOCANAR
Basketball	33.8	34.2	50	55.4
Body fat	44.2	46.4	84	47.2
Quake	43.8	46.4	44.87	28.2

Table 4. Comparison in terms of Confidence

Dataset	Alatas [52]	RPSO [53]	MOGAR [45]	MOCANAR
Basketball	0.60	0.60	0.83	0.82
Body fat	0.59	0.61	0.85	0.91
Quake	0.62	0.63	0.82	0.84

Table 5. Comparison in terms of Support

Dataset	Alatas [52]	RPSO [53]	MOGAR [45]	MOCANAR
Basketball	32.21	36.44	36.69	66.1
Body fat	63.29	65.22	65.26	79.57
Quake	38.74	38.74	36.96	51.22

Table 6. Spent Time to 15 runs

Dataset	Basketball	Body fat	Quake
Times	7m and 33 s	25m and 2s	11m and 21s

The average of the extracted rules length average in 10 runs is shown in Table 7. To compute Table 7, the average of the extracted rules length in each run is calculated and after completion of the runs, the average of the averages is calculated. In Table 8, the coverage values of the four algorithms on each dataset are shown. It shows that also in terms of coverage, our method is better than others. Finally, the results of the MOCANAR and the MOGAR algorithms are compared in terms of the interestingness and comprehensibility measures in Table 9. Because our support values are high, interestingness of the rules that are extracted with our method is low.

Table 7. Comparison in terms of Size

Dataset	Alatas [52]	RPSO [53]	MOGAR [45]	MOCANAR
Basketball	100.0	100.0	100.0	100.0
Body fat	84.12	86.11	93.52	99.
Quake	87.6	87.92	91.07	99.26

Table 8. Comparison in terms of Coverage

Dataset	Alatas [52]	RPSO [53]	MOGAR [45]	MOCANAR
Basketball	100.0	100.0	100.0	100.0
Body fat	84.12	86.11	93.52	99.
Quake	87.6	87.92	91.07	99.26

Table 9. Comparison in terms of Interestingness and Comprehensibly

Dataset	Interestingness		Comprehensibility	
	MOGAR	MOCANAR	MOGAR	MOCANAR
Basketball	0.53	0.38	0.72	0.92
Body fat	0.56	0.41	0.80	0.85
Quake	0.46	0.34	0.68	0.95

5. CONCLUSION

Because the extraction of association rules from numeric features has a very large search space, MOCANAR is suggested in the paper. Having high support, confidence, interesting and comprehensibility measures are the objectives that were considered in the MOCANAR. The rules were extracted incrementally in which, in the each increment of the algorithm, low numbers of high quality rules were made. Also in this paper, a comprehensive taxonomy of meta-heuristic algorithm was presented. Using this taxonomy, we decided to use Cuckoo Search algorithm because this algorithm is one of the most matured algorithms and also, it is a simple and understandable algorithm. In addition, until now this method was not used as a multi-objective algorithm and was not used in the association rule mining area. We demonstrate with our results that our method has high quality results in terms of our four objectives.

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, 1993, pp. 207-216.

[2] K. H. C. Lee, Y. Tse, G. Ho, K. Choy, and H. K. Chan, "Fuzzy association rule mining for fashion product development," Industrial Management & Data Systems, vol. 115, 2015.

[3] Ö. M. Soysal, "Association rule mining with mostly associated sequential patterns," Expert Systems with Applications, vol. 42, pp. 2582-2592, 2015.

[4] V. K. Ravi, P. Rahul, and S. K. Anand, "Designing an Expert System Using Association Rule Mining for Instant Business Intelligence," Middle-East Journal of Scientific Research, vol. 23, pp. 88-93, 2015.

[5] S. Kirkpatrick and M. Vecchi, "Optimization by simmulated annealing," science, vol. 220, pp. 671-680, 1983.

[6] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "GSA: a gravitational search algorithm," Information Sciences, vol. 179, pp. 2232-2248, 2009.

[7] N. Tayarani and M. Akbarzadeh-T, "Magnetic optimization algorithms a new synthesis," in Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on, 2008, pp. 2659-2664.

[8] S. Boettcher and A. G. Percus, "Extremal optimization: Methods derived from co-evolution," arXiv preprint math/9904056, 1999.

[9] Z. W. Geem, J. H. Kim, and G. Loganathan, "A new heuristic optimization algorithm: harmony search," Simulation, vol. 76, pp. 60-68, 2001.

[10] A. Eraser, "Simulation of genetic systems by automatic digital computers. I. Introduction," Australian Journal of Biological Sciences, vol. 10, pp. 484-491, 1957.

[11] J. R. Koza and P. James, "Rice, Genetic programming (videotape): the movie," ed: MIT Press, Cambridge, MA, 1992.

[12] I. Rechenberg, "Cybernetic solution path of an experimental problem," 1965.

[13] L. J. Fogel, A. J. Owens, and M. J. Walsh, "Artificial intelligence through simulated evolution," 1966.

[14] M. Dorigo, "Optimization, learning and natural algorithms," Ph. D. Thesis, Politecnico di Milano, Italy, 1992.

[15] M. S. Abadeh, J. Habibi, and E. Soroush, "Induction of Fuzzy Classification systems via evolutionary ACO-based algorithms," computer, vol. 35, p. 37, 2008.

[16] R. Hedayatzadeh, F. A. Salmassi, M. Keshtgari, R. Akbari, and K. Ziarati, "Termite colony optimization: A novel approach for optimizing continuous problems," in Electrical Engineering (ICEE), 2010 18th Iranian Conference on, 2010, pp. 553-558.

[17] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in Proceedings of the sixth international symposium on micro machine and human science, 1995, pp. 39-43.

[18] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition," in Evolutionary computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 4661-4667.

[19] X.-S. Yang, "Firefly algorithm, stochastic test functions and design optimisation," International Journal of Bio-Inspired Computation, vol. 2, pp. 78-84, 2010.

[20] M. M. Eusuff and K. E. Lansey, "Optimization of water distribution network design using the shuffled frog leaping algorithm," Journal of Water Resources Planning and Management, vol. 129, pp. 210-225, 2003.

[21] S.-C. Chu, P.-W. Tsai, and J.-S. Pan, "Cat swarm optimization," in PRICAI 2006: Trends in Artificial Intelligence, ed: Springer, 2006, pp. 854-858.

[22] W. Pan, "A new evolutionary computation approach: Fruit Fly Optimization Algorithm," in 2011 Conference of Digital Technology and Innovation Management, Taipei. Program code on the website http://www.oitecshop.byethost16.com/FOA. html, 2011.

[23] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," Control Systems, IEEE, vol. 22, pp. 52-67, 2002.

[24] X.-l. Li and J.-x. Qian, "Studies on Artificial Fish Swarm Optimization Algorithm based on Decomposition and Coordination Techniques [J]," Journal of Circuits and Systems, vol. 1, pp. 1-6, 2003.

[25] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in Nature inspired cooperative strategies for optimization (NICSO 2010), ed: Springer, 2010, pp. 65-74.

[26] B. Wang, X. Jin, and B. Cheng, "Lion pride optimizer: An optimization algorithm inspired by lion pride behavior," Science China Information Sciences, vol. 55, pp. 2369-2389, 2012.

[27] A. H. Gandomi and A. H. Alavi, "Krill herd: a new bio-inspired optimization algorithm," Communications in Nonlinear Science and Numerical Simulation, vol. 17, pp. 4831-4845, 2012.

[28] R. Oftadeh, M. Mahjoob, and M. Shariatpanahi, "A novel meta-heuristic optimization algorithm inspired by group hunting of animals: Hunting search," Computers & Mathematics with Applications, vol. 60, pp. 2087-2098, 2010.

[29] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, 2009, pp. 210-214.

[30] M. S. Abadeh, Evolutionary Algorithms and Biological Algorithms, 2012.

[31] Y. Chen, F. Li, and J. Fan, "Mining association rules in big data with NGEP," Cluster Computing, pp. 1-9, 2015.

[32] A. M. Palacios, J. L. Palacios, L. Sánchez, and J. Alcalá-Fdez, "Genetic learning of the membership functions for mining fuzzy association rules from low quality data," Information Sciences, vol. 295, pp. 358-378, 2015.

[33] F. Jiang, "Study on Adaptive Genetic Simulated Annealing Algorithm in Association Rules Mining," in Applied Mechanics and Materials, 2015, pp. 77-82.

[34] R. Kuo and C. Shih, "Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan," Computers & Mathematics with Applications, vol. 54, pp. 1303-1318, 2007.

[35] R. Kuo, S. Lin, and C. Shih, "Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan," Expert Systems with Applications, vol. 33, pp. 794-808, 2007.

[36] R. J. Kuo, C. M. Chao, and Y. Chiu, "Application of particle swarm optimization to association rule mining," Applied soft computing, vol. 11, pp. 326-336, 2011.

[37] A. Abraham and L. Jain, Evolutionary multiobjective optimization: Springer, 2005.

[38] M. S. Abadeh, J. Habibi, M. Daneshi, M. Jalali, and M. Khezrzadeh, "Intrusion detection using a hybridization of evolutionary fuzzy systems and artificial immune systems," in Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 3547-3553.

[39] J. D. Schaffer, "Multiple objective optimization with vector evaluated genetic algorithms," in Proceedings of the 1st International Conference on Genetic Algorithms, Pittsburgh, PA, USA, July 1985, 1985, pp. 93-100.

[40] Y. Li, X. Li, and J. N. Gupta, "Solving the multi-objective flowline manufacturing cell scheduling problem by hybrid harmony search," Expert Systems with Applications, vol. 42, pp. 1409-1417, 2015.

[41] J. Rudy and D. Żelazny, "Solving multi-objective job shop problem using nature-based algorithms: new Pareto approximation features," An International Journal of Optimization and Control: Theories & Applications (IJOCTA), vol. 5, pp. 1-11, 2014.

[42] K. Gao, P. Suganthan, Q. Pan, T. Chua, T. Cai, and C. Chong, "Pareto-based grouping discrete harmony search algorithm for multi-objective flexible job shop scheduling," Information Sciences, vol. 289, pp. 76-90, 2014.

[43] J. Gómez, C. Gil, R. Baños, A. L. Márquez, F. G. Montoya, and M. Montoya, "A Pareto-based multi-objective evolutionary algorithm for automatic rule generation in network intrusion detection systems," Soft Computing, vol. 17, pp. 255-263, 2013.

[44] P. P. Wakabi-Waiswa and V. Baryamureeba, "Extraction of interesting association rules using genetic algorithms," International Journal of Computing and ICT Research, vol. 2, pp. 26-33, 2008.

[45] B. Minaei-Bidgoli, R. Barmaki, and M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms," Information Sciences, vol. 233, pp. 15-24, 2013.

[46] S. Walton, O. Hassan, K. Morgan, and M. Brown, "Modified cuckoo search: a new gradient free optimisation algorithm," Chaos, Solitons & Fractals, vol. 44, pp. 710-718, 2011.

[47] N. Bacanin, "Implementation and performance of an object-oriented software system for cuckoo search algorithm," International Journal of Mathematics and Computers in Simulation, vol. 6, pp. 185-193, 2012.

[48] R. Rajabioun, "Cuckoo optimization algorithm," Applied soft computing, vol. 11, pp. 5508-5518, 2011.

[49] A. H. Gandomi, X.-S. Yang, and A. H. Alavi, "Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems," Engineering with computers, vol. 29, pp. 17-35, 2013.

[50] X.-S. Yang and S. Deb, "Cuckoo search: recent advances and applications," Neural Computing and Applications, vol. 24, pp. 169-174, 2014.

[51] H. A. Guvenir and I. Uysal, "Bilkent university function approximation repository," ed, 2000.

[52] B. Alataş and E. Akin, "An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules," Soft Computing, vol. 10, pp. 230-237, 2006.

[53] B. Alatas and E. Akin, "Rough particle swarm optimization and its applications in data mining," Soft Computing, vol. 12, pp. 1205-1218, 2008.

AUTHORS

Irene Kahvazadeh was born on September 1991. She passed her bachelor between 2009 to 2012 years. From February 2013, she is student in master degree at Tarbiat Modares University.

Mohammad Saniee Abadeh received his B.S. degree in Computer Engineering from Isfahan University of Technology, Isfahan, Iran, in 2001, the M.S. degree in Artificial Intelligence from Iran University of Science and Technology, Tehran, Iran, in 2003 and his Ph.D. degree in Artificial Intelligence at the Department of Computer Engineering in Sharif University of Technology, Tehran, Iran in February 2008. Dr. Saniee Abadeh is currently a faculty member at the Faculty of Electrical and Computer Engineering at Tarbiat Modares University. His research has focused on developing advanced meta-heuristic algorithms for data mining and knowledge discovery purposes. His interests include data mining, bio-inspired computing, computational intelligence, evolutionary algorithms, fuzzy genetic systems and memetic algorithms.

INTENTIONAL BLANK

ASSOCIATION RULE DISCOVERY FOR STUDENT PERFORMANCE PREDICTION USING METAHEURISTIC ALGORITHMS

Roghayeh Saneifar and Mohammad Saniee Abadeh

Faculty of Electrical and Computer Engineering, Tarbiat Modares University,
Tehran, Iran
r.saneifar@modares.ac.ir
saniee@modares.ac.ir

ABSTRACT

According to the increase of using data mining techniques in improving educational systems operations, Educational Data Mining has been introduced as a new and fast growing research area. Educational Data Mining aims to analyze data in educational environments in order to solve educational research problems. In this paper a new associative classification technique has been proposed to predict students final performance. Despite of several machine learning approaches such as ANNs, SVMs, etc. associative classifiers maintain interpretability along with high accuracy. In this research work, we have employed Honeybee Colony Optimization and Particle Swarm Optimization to extract association rule for student performance prediction as a multi-objective classification problem. Results indicate that the proposed swarm based algorithm outperforms well-known classification techniques on student performance prediction classification problem.

KEYWORDS

Educational data mining, bee colony optimization, continuses rule extraction, classification, particle swarm optimization

1. INTRODUCTION

As the volume of archived data increases, the need for more efficient and faster data analysis techniques increases concurrently. All of the saved records in databases of organizations would be useless, if decision makers do not employ effective knowledge discovery techniques. Data mining methods analyze huge amount of databases to discover valuable and ready to use knowledge [8].

Nowadays, data mining techniques have been used in academic and educational environments and leave a remarkable effect in this domain [9]. Educational Data Mining (EDM) refers to the employment of knowledge discovery techniques and methods in education. The main goal of EDM is to enhance various educational activities such as student performance prediction, education facility improvement, etc.

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 115–123, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51510

As mentioned above, EDM is a domain that uses machine learning, data mining and statistical techniques, analyses educational data. Thanks to employ of these techniques, it is possible to improve the learning/teaching processes involving students or instructors.

Educational data come in many different and very complex formats. The last surveys in this scope is related to (Alejandro Pena-Ayala,2013), establishing the following EDM approaches [1]:

- Student behavior modeling
- Student performance modeling
- Student modeling
- Assessment
- Curriculum, domain knowledge, sequencing, and teachers support
- Student support and feedback

Other survey is related to Romero and Ventura [2], which is survey on educational data mining between 1995 and 2005. Using data mining techniques in higher education is a recent research domain; there are a lot of works in this area. That is because of its potentials to educational institutes.

Ayesha et al. employed the k-means data mining clustering algorithm to predict students' learning activities in an educational database including classroom quizzes, final and mid exam and other assignments. This correlated information will be conveyed to the teacher before the transfer of final exam. This study helps the teachers to improve the performance of students and reduce the failing ratio by taking appropriate steps at on time [3].

Baradwaj and Pal, in the year 2011, used the classification as data mining methods to evaluate student' performance, they applied decision tree technique for classification. The aim of their research is to extract knowledge that describes students' performance in end semester quizzes. They used students' educational data from the student' previous database including Class test , Assignment marks , Attendance, , Seminar. This study helps sooner in identifying the students who need more attention and allow the teacher to provide appropriate advising [4].

Chandra and Nandhini, applied the association rule mining method based on students courses to identifies students' break patterns. The aim of their research is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances. The extracted association rules lay out some hidden patterns of students' courses which could serve as a foundation stone for academic planners in making decisions and modification and an aid in the curriculum re-structuring with a view to improving students' performance and reducing break rate [5].

Shannaq et al, used the classification since data mining technique to predict the numbers of listed students by evaluating academic data from enrolled students to study the main attributes that may affect the students' truth (number of enrolled students) [6].The decision tree as a classification method to extract classification rules and the extracted classification rules are analyzed and evaluated using different evaluation methods. It allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to focus over in higher education systems for support and feedback.

Made a prediction model using the GP method to identify at-risk students in traditional school settings. A feature selection technique was used to reduce the attributes [7].

Wolff et al. (2013) have applied a decision-tree as data mining techniques to identify at-risk students in a virtual learning environment.

In this paper a new associative classification technique has been proposed to predict students final performance. In this research work, we have employed Honeybee Colony Optimization and Particle Swarm Optimization to extract association rule for student performance prediction as a multi-objective classification problem. Results indicate that the proposed swarm based algorithm outperforms well-known classification techniques on student performance prediction classification problem.

The rest of this paper is organized as follows: Section 2 presents the new proposed classification method for student performance prediction.

2. PROPOSED METHOD

In this section, we introduce a new approach, called Bee-RM, of multi-objective optimization based on the optimization of bee colony algorithm and particle swarm optimization.

In the following, we present the outlines of our proposed approach.

Association rule extraction is widely used data mining tasks. This is due to the interpretability feature of these rules for non-experts. The extraction of the association rules is usually performed using the meta-heuristic algorithms. In this paper, we take two major factors into consideration regarding the classification: the first one is the accuracy and the second is Interpretability.

The knowledge base used in this work is presented as a rule base. It is an important issue to select a set of optimum rules in these systems. In our Bee-RM approach, the rule extraction is performed using "pareto optimality" and considering the multi-objective factor.

Since there is rarely a unique solution which optimizes all objective functions, we look for a trade-off between objectives instead of seeking a unique solution for multi objective optimization.

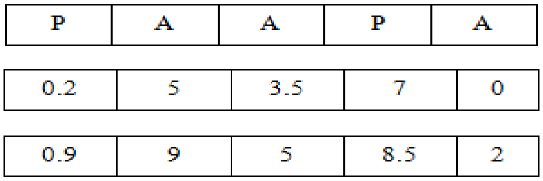
2.1 RULE GENERATION BY BEE_RM

In this work, we decided to continuously extract rules as there is only few works which perform continuous rule extraction. The advantage of the continuous rule extraction is that the whole space is explored. However, the whole space exploration needs a lot of space, which demands to use more powerful algorithms.

In the following, we present how to model the association rules using the bee colony optimization and particle swarm optimization (PSO). Each member of the population is presented as an array with three rows. Then, each association rule is created by a member.

Since rules are created for each class, we use class zero as an example.

In the first array, “A” presents absence and “P” presents the specific property in the rule. In this approach, we do not need to perform bins and so the span is seen continuously.



P	A	A	P	A

0.2	5	3.5	7	0

0.9	9	5	8.5	2

The second array’s values show the lower limit of each property. The third array shows the upper limit of each property. Therefore, the rule presented by these arrays is:

If (0.2<F1<0.9 and 7<F4<8.5) then class = 0

The first array contains discrete values, in the ConstructSolution function, we use the bee colony optimization in order to predicate and in the case of two other arrays which present the span, we use PSO optimization.

In the first fold of each category in the dataset, the generation is performed “MaxGeneration “ times. Inside each generation, the population size is equal to the value of “Population” parameter. In each execution of the algorithm, for each class in the dataset, the generation is performed and every member of the population produces the optimized results. Then, we use the “optimized association rules extracted for all classes” as input of the classifier method in order to classify the test dataset.

Finally, the average accuracy obtained by 10-fold execution is considered as the main accuracy of the Bee_RM algorithm.

2.2 HONEYBEE HIVE OPTIMIZATION (HHO)

The “ConstructSolution” method for optimizing the first array, create a path for each bee according to the Dance Table and heuristic information. (1)

$$P_k(r,s) = \begin{cases} \frac{[\delta(r,s)]^{\alpha}[\eta(r,s)]^{\beta}}{\sum_{u \in J_{k\circledR}}[\delta(r,s)]^{\alpha}[\eta(r,s)]^{\beta}} & \text{if } s \in J_k \\ 0 & \end{cases} \tag{1}$$

The original fitness function, presented in this paper, is implemented according to the Eq (2). Below we demonstrate this function in (2).

$$H(r,s) = p_1 \times \text{support(solution)} + p_2 \times \frac{\#\text{non Don'tcare}}{\#\text{features}} \tag{2}$$

In this formula (2) P_1 is the effectiveness and importance given to the support of produced solution, and P2 is the importance given to the “Don’t-care” relative to the number of all features.

2.3 PARTICLE SWARM OPTIMIZATION (PSO)

We use the particle swarm optimization (PSO) algorithm in a continuous space and in multi objective form. The objective function in Eq (3) is used to calculate the Local-best found by each individual inside the same individual. The Global-best found in the whole population of individuals is kept in another variable called Gbest in each individual. In other word, we do not have a unique Global-best but many.

Fitness = Support Percent*Support (solution) + (1-SupportPercent)* Confidence (solution) (3)

All Gbest are the most optimized local Non-dominated association rules obtained by Eq (4) optimization in the current population. To calculate the location of the next move of particle, we use the average of these local Non-dominated rules as demonstrated in the Eq(5) and Eq(6).

$$\propto = \frac{\sum_{i=1}^{n} \mathrm{Gbest}_i - x_i(t)}{n} \qquad \text{n= Local Non Dominated rules} \tag{4}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{5}$$

$$v_i(t+1) = c_1 r_1 \left(\mathrm{Lbest} - x_i(t)\right) + c_2 r_2 \propto \tag{6}$$

The more general rules cover a big span of the dataset records. It reduces the interestingness of the rules. Our objective is to make a trade-off between interestingness and support value of obtained association rules. We try to extract more detailed association rules with high support value and interestingness by defining the “Interval-p” parameter.

2.4 STOPPING CONDITION

Once all rules are created by all members of the current population, local non-dominated and global non-dominated rules are determined. The most important condition to stop the training phase is a constant number of repetitions. The members continue the procedure till the stop condition is satisfied. The procedure stops if the repetition number of procedure is reached (the “Maxgeneration” number). Then, the best association rules according the Pareto-optimality optimization are selected.

3. EXPERIMENTAL RESULTS

This section shows the experimental results of the proposed method versus other classification techniques. Our proposed method will be analyses educational data generated on a Moodle platform.

Moodle’s log is the baseline system used in this research. Moodle is a free virtual learning environment (VLE). Moodle is therefore evolving system and dynamic. Anyone can download and install it. An administrator is responsible for managing users (students, teachers, etc.) and course virtual classrooms. The Moodle system view differs depending on the role the user plays (teacher, student, administrator etc.).

Moodle is developed by programmers as an open source system, from all over the world. As of 2013, Moodle system has over 77,000 registered sites in over 215 countries. It prepares support to over 65 million students all over the world, trained by over 1.2 million teachers.

Moodle is only one of many support tools for virtual learning environment (VLE). There are other similar distance systems like, for example, ATutor, eCollege, Desire2Learn or Dokeos.
The information of interaction is stored as attributes in a user (student) profile. In our data set, 11 attributes and values are stored, with 357 records. These attributes include: number of interaction between student-student, student-teacher, and etc. Detail of this data set is as follows. Table 1 shows detail information about attributes of Moodle data set.

Experimentally, we have tried to set the best parameters for proposed method. The values of different user-defined parameters of Bee_RM is reported in Table 2.

Table 1.Information about features of Moodle dataset.

Category	**features**
Category 1 Based on agent	ST-ST :Student – student ST-TE: student –teacher ST-CO :Student – content ST-SY : Student-system
Category 2 Based on frequency of use	TC :Transmission of contents CI: Creating class interactions SA :Student assessment / evaluating students
Category 3 Based on participation mode	AC : Active PA: Passive
Dependent variable - Academic performance	GR: Final grade

The performance of Bee_RM is evaluated using 10-fold cross-validation test (Michalski et al., 1998). In this section of research, the all obtained results are reported. Important scale to evaluate the proposed method : accuracy.

The accuracy is the number of instances correctly classified and being calculated according to Eq. (7)

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{7}$$

Table 2. Parameter setting of Bee_RM.

Parameter	Value
$\text{Population}_{\text{Size}}$	30
Maxgeneration	150
DefultDancers	6
C	0.5 , 0.03
P	1 , 4
SupportPercent	0.5
Interval_p	0.5
α, β	2 , 1

Figs.1 and 2 denote the effect of different population sizes of the new proposed metaheuristic algorithm on accuracy and execution time respectively. Fig. 3 shows the Influence of P_2 parameter on average length of rules.

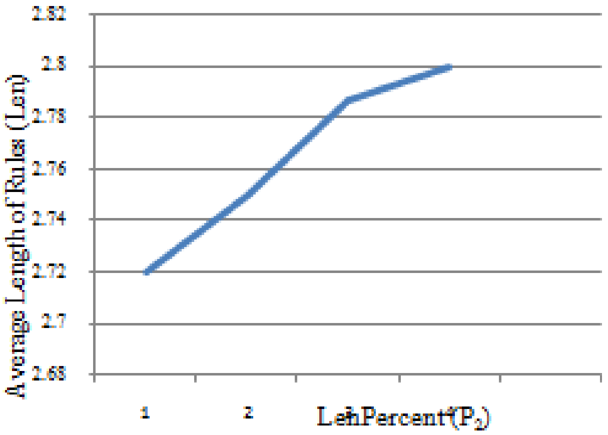
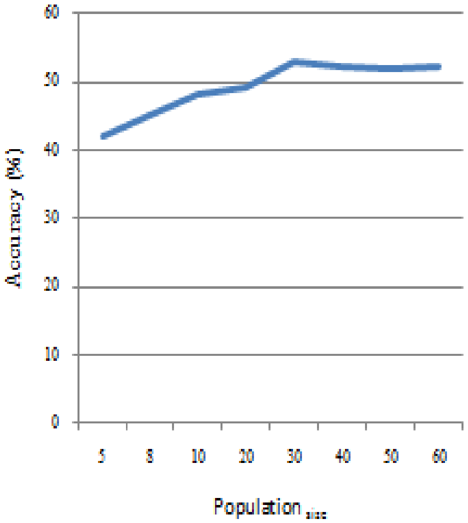


Figure 3. Infuluence of P_2 parameter on average length of rules.



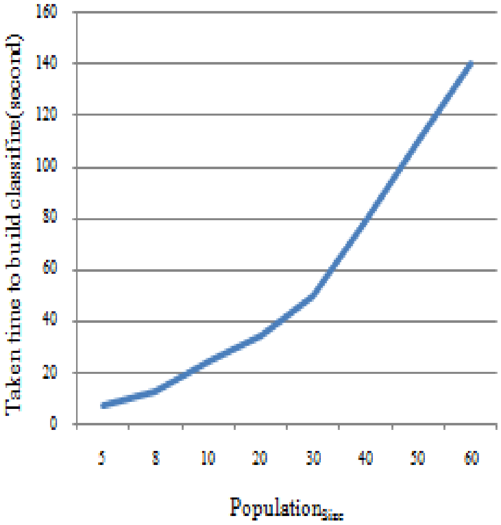


Figure 1. Influence of number of individual on accuracy

Fig 2. Influence of number of individual on taken time to learn the classifier

Table 3. Classification accuracy obtained with different method for Moodle.

Method	**Classification Accuracy (%)**	**Study**
KNN	47.29% +/- 6.05%	Cover & Hart, (1967) and Rapidminer tool is available
NN	51.68% +/- 3.83%	Nsky, (1954) and Rapidminer tool is available
Baysian	43.71% +/- 8.26%	Russell, Stuart, 1995) and Rapidminer tool is available
Rule Induction	46.63% +/- 6.55%	J. Stefanowski, (1998) and Rapidminer tool is available
PART	51.26	Witten and Frank, (2005) and WEKA tool is available
OneR	45.93	Weka: http://www.cs.waikato.ac.nz/~ml/weka/
JRip	50.42	Weka: http://www.cs.waikato.ac.nz/~ml/weka/
ZeroR	41.73	Witten and Frank, (2005) and WEKA tool is available
IBK	40.33	Witten and Frank, (2005) and WEKA tool is available
Logistic	46.49	Witten and Frank, (2005) and WEKA tool is available
SimpleLogistic	51.26	Witten and Frank, (2005) and WEKA tool is available
SMO	52.10	Witten and Frank, (2005) and WEKA tool is available
NaiveBayes	36.13\	Witten and Frank, (2005) and WEKA tool is available
ClassificationVia Regression	52.66	Witten and Frank, (2005) and WEKA tool is available
Vote	41.73	Witten and Frank, (2005) and WEKA tool is available
Random Tree	45.93	Weka: http://www.cs.waikato.ac.nz/~ml/weka/
Random Forest	47.05	Witten and Frank, (2005) and WEKA tool is available
J48	46.21	J.R. Quinlan, (1993) and WEKA tool is available
CPSO-C	42	Liu et al. 2004, and KEEL tool is available
SLAVEC	51	González and Pérez 2001, and KEEL tool is available
MPLCS-C	47	Bacardit and Krasnogor 2009, KEEL tool is available
C-SVM-C	51	KEEL tool is available
XCS-c	47	Wilson 1995, and KEEL tool is available
GFS-SP-C	48	Sánchez et al. 2001, and KEEL tool is available
Bee_RM	53.46%+/-5.46%	Our study

Table 3 shows accuracy of Bee_RM versus several recent and famous classification methods. We used 3 famous tools in data mining, for comparison.

To compare our results with other studies, we have used WEKA, Rapidminer and KEEL softwares.

Six evolutionary rule learning algorithms are used in which 3 of them learn fuzzy rules and 3 of them learn crisp rules in an evolutionary way. These results reveals, our proposed method Bee_RM using 10-fold cross validation obtains the highest classification accuracy, 53.46%, reported so far. So, we can draw this conclusion that the combination of Bee Colony Optimization and particle swarm optimization with continues logic, would be very effective in predicting student final performance in educational data.

Although there is not any accurate definition for interpretability of classification methods but the number of rules (NR) and mean length of rules(Len) are often mentioned as two main factors of interpretability.

4. CONCLUSIONS

In this paper we employed the capability of swarm based techniques to extract association rules for student performance prediction as a multi-objective classification problem. The proposed algorithm had a low convergence time and it used a few number of parameters. Honeybee Colony Optimization and Particle Swarm Optimization were the two used metaheuristics to extract association rules. The fitness function in both of these algorithms considers support and length of the association rules. Results showed that using the proposed metaheuristic-based rule discovery approach enables us to extract accurate and interpretable knowledge for student performance prediction. Our future works focus on using new proposed metaheuristic algorithms such as Gravity Search and Vortex Search Algorithm instead of PSO and Honeybee Colony. Moreover, we aim to consider other measures such as confidence, correlation and interestingness along with support and rule length.

REFERENCES

[1] Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." Expert systems with applications 41.4 (2014): 1432-1462.

[2] Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146

[3] Baradwaj, B. and Pal, S. (2011) 'Mining Educational Data to Analyze Student s' Performance', International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[4] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol.

[5] Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) 'Data Mining Model for Higher Education System', European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[6] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59. 47, no. 1, pp. 156-163.

[7] Marquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining

[8] Pieter, Adriaans. DolfZantinge, 1996. Data Mining (New York: Addison Wesley)

[9] D. T. Larose, Discovering knowledge in data: an introduction to data mining. Wiley.com, 2005.

INTENTIONAL BLANK

DATA AGGREGATION IN WIRELESS SENSOR NETWORK BASED ON DYNAMIC FUZZY CLUSTERING

Arezoo Abasi and Hedieh Sajedi

Department of Mathematics, Statistics and Computer Science,
College of Science, University of Tehran, Tehran, Iran
arezoo_abasi@ut.ac.ir
hhsajedi@ut.ac.ir

ABSTRACT

Wireless Sensor Networks (WSN) use a plurality of sensor nodes that unceasingly collected and sent data from a specific area to a base station. Cluster based data aggregation is one of the popular protocols in WSN. Clustering is an important procedure for extending the network lifetime in WSNs. Cluster Heads (CH) aggregate data from relevant cluster nodes and send it to the base station. A main challenge in WSNs is to select suitable CHs. In another communication protocol based on a tree construction, energy consumption is low because there are short paths between the sensors. In this paper, we propose Dynamic Fuzzy Clustering (DFC) data aggregation. The proposed method first uses fuzzy decision making approach for the selection of CHs and then a minimum spanning tree is constructed based on CHs. CHs are selected efficiently and accurately. The combining clustering and tree structure is reclaiming the advantages of the previous structures. Our method is compared to Low Energy Adaptive Clustering Hierarchy (LEACH), Cluster and Tree Dara Aggregation (CTDA), Modified Cluster based and Tree based Data Aggregation (MCTDA) and Cluster based and Tree based Power Efficient Data Collection and Aggregation (CTPEDCA).Our method decreases energy consumption of each node. In DFC data aggregation, the node lifetime is increased and the survival of the WSN is improved.

KEYWORDS

Wireless sensor networks; Data aggregation; Clustering; Minimum Spanning Tree; Fuzzy decision making.

1. INTRODUCTION

In WSN, sensor nodes are usually scattered randomly in large numbers. In this area, there is no opportunity for maintenance and battery replacement for the most of the applications, which use the sensor nodes to surveillance the remote field [1].

The sensor's battery is limited. The lifetime on each node depends on the power that has significantly affected the relationship between the nodes. One of the accurate requirements of these nodes is the efficient use of the saved energy. Multiple algorithms have been designed for

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 125–138, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51511

impressive handling of nodes energy in WSNs using several clustering schemes [2, 3]. Optimal data aggregation can save nodes energy. In this sensor network data are gathered by the sensor nodes from our study area. There is a data transmission method that merges data from several sensor nodes into one pack which is data aggregation. Decreasing the disjointed communication at different levels and in turn to reduce the total energy consumption is the main aim of data aggregation. There are dissipated different amounts of energy to process raw data. There are two popular protocols: Cluster based data aggregation [4] and Tree based data aggregation [5]. Some of WSNs consists of clusters, in which each cluster has a CH. CHs have a significant impress in network lifetime. An ideal CH is the one which has the highest residual energy, maximum number of neighbor nodes around the CH and the shortest distance from the base station [6]. Whatever the selected CH is more similar to the ideal CH, network lifetime is increased.

We can use Multiple Attribute Decision Making (MADM) approach to select CHs with multi criteria [7]. This method selects alternatives based on their multiple criteria. The main problem is the difficult estimation of the exact values of all the criteria. Synchronous consideration of all criteria in CHs selections can be used MADM approach. In case of multi criteria fuzzy based MADM methodologies are efficient and impressive [8, 9].

In this paper, we proposed a hybrid approach called DFC data aggregation, which gathers and combines data and avoids redundant data transformations, therefore successively saves energy and bandwidth.

Proposing DFC data aggregation, we preserve the advantages and minimize the disadvantages of the clustering and tree based approaches. We use DFC data aggregation to extend the lifetime of WSNs and energy consumption of sensor nodes. The optimized CHs are selected to spread energy efficiently using multi criteria. CHs are selected based on the residual energy, the number of neighbor nodes and distance from the base station. After cluster formation, CHs receive data from member nodes in clusters, aggregate data and send it to the base station. A spanning tree covers all the sides as vertices and consists no cycles. The tree is constructed in the procedure that the node with the smallest identifier is chosen as the root [10, 11]. All the nodes with the shortest path conjunct to the selected root. The protocol requires that each node exchanges configuration messages in a specific format which contains its own identifier, its chosen root, and the distance to this selected root. Each node updates its configuration message upon identifying a root with a smaller identifier or the shortest-path neighbor. In addition, the neighbor for which the shortest route configuration message comes from is chosen as the parent of a node.

In this paper, we employ multi criteria decision making approach, Fuzzy Analytic Hierarchy Process (FAHP) and hierarchical fuzzy in clusters on WSNs [12, 13]. AHP considers a set of assessment criteria, and a set of alternative choices among which the best decision is to be made. AHP generates a weight for each evaluation criteria according to the comparisons of the criteria. The superior the weight, the more significant the corresponding criterion. The AHP method could improve the network lifetime.

In this research, we also analyze LEACH [14], CTDA [15], MCTDA [16], and CTPEDCA [17]. We compared our proposed method with these methods in terms of energy consumption and the amount of energy remaining in each sensor network lifetime. Simulation conclusions illustrate that our proposed approach is more efficient than other algorithms.

2. RELATED WORKS

Clustering in WSNs is an effective procedure to decrease the energy consumption of sensor nodes. In cluster based routing algorithms for wireless networks, LEACH is famous because it is simple and efficient. In LEACH, CH nodes are selected randomly and all the non CH nodes are formed based on the received signal power from the CHs. In LEACH each node can become a CH, there is no pattern in electing CHs and all nodes have the same chance to be a CH, thus LEACH is not efficient. CHs are selected randomly and the energy is divided between all the nodes equally. CHs aggregate all received data from all nodes in the clusters [14].

LEACH forms clusters based on the received signal strength and uses the CHs as portals to the sink. All the data processing like data fusion and aggregation are locally accomplished into the cluster. CH is selected periodically among the nodes of the cluster. LEACH forms distributed clusters, where nodes make independent decisions without any concentrated control. In LEACH, each CH has a straight communicates with the base station no matter the distance is close or not. When the network is massive, the communication between CHs and the base station consumes much energy for the long distance transmission. In LEACH, the size of clusters can be increased if the number of CHs is reduced. This makes induced excessive delays introduced by the number of nodes in the same cluster [18, 19].

CTDA is a hybrid cluster and tree based algorithm and is proposed for data aggregation. It employs a data aggregation mechanism in the CH to lessen the amount of data transmitted. Therefore, CTDA decreases the energy dissipation in communication. CTDA decreases data transfer volume so it enhances energy efficiency and attains the purpose of saving energy of the sensor nodes. CTDA decreases the number of nodes, which directly send data to the base station. In WSN with constrained energy, it is inefficient for sensors to select CHs randomly. CTDA method does not perform any calculation in choosing the CHs and select CHs randomly. It is non optimal to selected CHs by chance because it imposes an additional burden to the network. CTDA does not consider the amount of remaining energy in the nodes and it increases the wasted energy and decreases the lifetime of the network [15].

In MCTDA method, minimum spanning tree does not do data aggregation and only data of CHs by tree structure is sent to the base station [16].

CTPEDCA uses the full distribution in hierarchical WSNs. CTPEDCA is based on clustering and Minimum Spanning Tree routing strategy for CHs and the time complexity is small. CTPEDCA can balance the energy consumption of all the nodes, particularly the CH nodes in each round and extend the lifetime of the networks. In each round, CTPEDCA allows only one CH communicate directly to the base station. In CTPEDCA, a CH with the maximum remaining energy is selected as the base, CH0. CH0 constructs a minimum spanning tree between all CHs and broadcasts tree information for all the CHs. If the number of CH is K, K-1 CHs send data only to CH0 and CH0 transmit data to the base station. The disadvantage of this method is the network is dependent on the CH0. CH0 is placed under pressure and needs a lot of energy. If CH0 is failed, the network also failed. When the base station is too far, this method is not useful [17].

In WSN, improving the energy performance and maximizing the networking lifetime are the main challenges. For this reason a hierarchical clustering scheme, called Location Energy Spectral Cluster Algorithm (LESCA) is proposed in [20]. LESCA specifies the number of clusters in a WSN automatically. It is based on spectral classification and considers both the remaining energy and some properties of nodes. LESCA uses the K-way algorithm and proposes new features of the network nodes such as average energy, distance to the base station, and distance to cluster centers in order to determine the clusters and to elect the CHs of a WSN. If the clusters are not constructed in an optimal way and/or the number of the clusters is greater or less than the optimal number of clusters, the total consumed energy of the sensor network per round is increased exponentially.

3. ASSUMPTION

We consider the following assumption:

- All the nodes know their location and nodes are distributed randomly in the experimental area.
- The base station has no energy constraint and is located at the top of the area.
- The initial number of CHs is constant and does not change over time.

The superiority of protocols is changed because there are different presumptions about the radio features, such as energy dissipation in transmitter and receiver models. In our plan, a simple model is used for the radio energy dissipation which is the transmitter, power amplifier, and receiver dissipates energy to run the radio electronics [21]. The distance between the transmitter and the receiver is used for the free space (d^2 power loss) and the multipath fading (d^4 power loss) channel models.

In general, the free space (*fs*) model is used when the distance is less than a threshold d0 and if more than the threshold d0, the multipath (*mp*) model is used [21]. Therefore, when n bit data message is transmitted over a distance d to achieve an acceptable signal, the energy expended by the radio E_{TX} can be expressed as Eq.(1).

$$E_{TX}(n,d) = \begin{cases} n\, E_{elec} + n\, \varepsilon_{fs}\, d^2 & d \le d_0 \\ n\, E_{elec} + n\, \varepsilon_{mp}\, d^4 & d \ge d_0 \end{cases} \qquad (1)$$

where, ε_{fs} is the energy consumed by the amplifier to transmit at a shorter distance. ε_{mp} is the energy consumed by the amplifier to transmit at a longer distance. E_{elec} is the energy dissipated in the electronic circuit to transmit or receive the signal, which relied on agents such as the digital coding, modulation, filtering and spreading of the signal. E_{RX} is the radio energy consumed to receive this message, which is calculated by Eq.(2).

$$E_{RX}(n) = n * E_{elec} \qquad (2)$$

4. THE PROPOSED ALGORITHM

In this paper, we propose an algorithm for data aggregation called Dynamic Fuzzy Clustering (DFC) data aggregation. DFC data aggregation uses the concepts of cluster and tree based algorithms. The main idea of the cluster based routing is to lessen the amount of data transmission via engage the data aggregation mechanism in the CH. DFC data aggregation decreases the energy dissipation and saves the residual energy of the nodes. DFC data aggregation has three phases:

Phase 1. CHs selection
Phase 2. Cluster construction
Phase 3 Tree formation of CHs

Our proposed method is inspired from two approaches named Pareto Optimal Solutions [22-23] and Fuzzy TOPSIS. At the beginning of the network, we select CHs based on Fuzzy TOPSIS [6]. The clusters are formed based on the distance between nodes and CHs. Then, the tree is organized due to CHs situation. This process continues until the first CH dies or the CH energy gets lower than a defined threshold. In this case, we determine CHs based on Fuzzy TOPSIS again. We determine the CHs based on maximizing the amount of energy efficiency. Although the initial number of CHs is assumed constant, but it can be dependent on several parameters, i.e., network topology, residual energy of nodes, and the relative costs of calculation versus communication. The iteration of the above mentioned steps creates rounds in our proposed algorithm. In the sequel, we describe the phases of DFC data aggregation.

4.1. CH Selection (Phase 1)

Multi Criteria Decision Making (MCDM) techniques have been applied to quantitative decision making problems [24]. MCDM can be divided into two main categories. Multiattribute decision making (MADM) approach [24] is one of the main categories of MCDM techniques. On the other hand multi objective decision making (MODM) [22] is another main category in MCDM techniques. In this paper, we use MODM (Pareto optimal technique) and MADM (fuzzy TOPSIS) for selecting CHs.

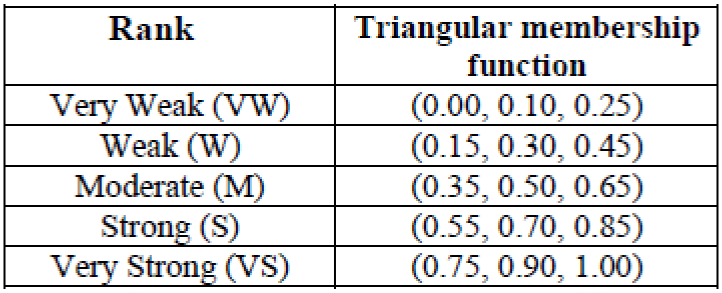
4.1.1. MODM (Pareto optimal technique)

The Pareto optimal solutions introduced by the economist Vilfredo Pareto [23]. Pareto technique determines the solution space which solutions are non dominated. Pareto solution space specifies an area which comprising of all conceivable solutions in multi objective decision making problems. The solution space is classified into three groups, namely, completely dominated, neither dominant nor dominating and non dominated.

4.1.2. MADM (Fuzzy TOPSIS)

It is often difficult to determine the exact values of attributes of the sensor nodes [6]. Thus, we use a fuzzy approach to determine the comparative significance of criteria instead of exact values. In this algorithm, five fuzzy linguistic variables are considered as the following: Very Weak, Weak, Moderate, Strong, and Very Strong. Figure 1 illustrates the fuzzy triangular functions. The triangular membership functions are determined in Table 1.

Table 1. Transformation of fuzzy triangular membership function



Rank	Triangular membership function
Very Weak (VW)	(0.00, 0.10, 0.25)
Weak (W)	(0.15, 0.30, 0.45)
Moderate (M)	(0.35, 0.50, 0.65)
Strong (S)	(0.55, 0.70, 0.85)
Very Strong (VS)	(0.75, 0.90, 1.00)

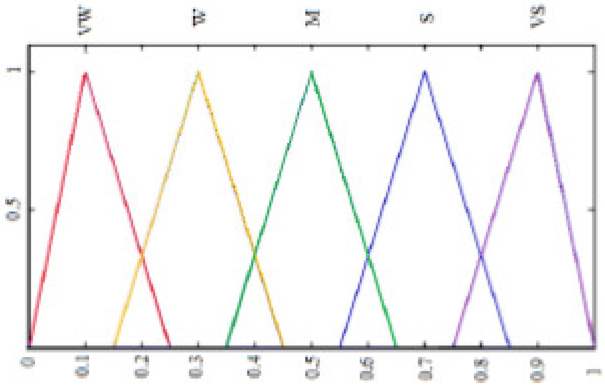


Figure 1. Fuzzy triangular function

In fuzzy TOPSIS approach, decision matrix has “m” alternatives and “n” attributes that could be assumed to be a problem of “n” dimensional hyper plane has “m” points whose location is given by the value of their attributes [8]. i and j are i=1,2,…,m and j=1,2,…,n. The decision matrix is as the following:

$$A = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & x_{ij} & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \qquad (3)$$

The weight of the jth column of matrix A is shown as (4):

$$C = \left[c_{1\,,}c_2, \cdots, c_j \cdots c_n\right] \qquad (4)$$

where x_{ij} and c_j are fuzzy numbers. We have determined 0.5, 0.25, and 0.25 weights to the remaining energy, number of neighbors, and distance from the sink, respectively. P is a fuzzy decision matrix which is normalized as the follow:

$$P = [p_{ij}]_{m\times n}$$

F is the weighted normalized fuzzy decision matrix.

$$F = \begin{bmatrix} c_1p_{11} & c_2p_{12} & \cdots & c_np_{1n} \\ c_1p_{21} & c_2p_{22} & \cdots & c_np_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ c_1p_{m1} & c_2p_{m2} & \cdots & c_np_{mn} \end{bmatrix} \qquad (5)$$

In order to simplify the above matrix (f_{mn}=c_np_{mn}), we summarize it as follows:

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{12} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix} \qquad (6)$$

The best conceivable solution is the shortest distance from the ideal solution, and the worst conceivable solution is the furthest distance from the ideal solution. The best and the worst solutions are obtained from the weighted normalized fuzzy decision matrix given by (6). The best solutions are denoted bt BS_j and WS_j denotes the worst Solutions:

$$BS_j = \{(\max f_{ij} \mid i = 1, 2, \ldots, m), j = 1,2, \ldots, n \} \qquad (7)$$

The worst solutions are defined as:

$$WS_j = \{(\min f_{ij} \mid i = 1, 2, \ldots, m), j = 1,2, \ldots, n \} \qquad (8)$$

We select a solution which is the nearest from the best conceivable solution and the furthest from the worst ideal solution. The distances of each alternative from the best solution and the worst solution are the separation measures. Distance of Best Solutions (DBS) and Distance of Worst Solutions (DWS) are as:

$$DBS_i = \sum_{j=1}^{n} d(f_{ij}, BS_j) \qquad i = 1,2, \ldots, m \qquad (9)$$
$$DWS_i = \sum_{j=1}^{n} d(f_{ij}, WS_j) \qquad i = 1,2, \ldots, m \qquad (10)$$

Rank indices of TOPSIS are estimated as:

$$Rank_i = \frac{DBS_i}{DWS_i + DBS_i} \qquad (11)$$

Superior TOPSIS rank nodes are selected as the CHs. Each selected CH gets a unique identifier (ID).

4.2. Cluster Construction (Phase 2)

All the selected CHs disseminated identity message to non CH nodes in the network. Each node calculates the distance from all the CHs then joins to the cluster, which has the minimum distance from its CH. K specifies the number of CHs. A distance matrix is used for reclustring nodes based on the distance to the selected CHs. The distance metric used here is the Euclidean metric. The Euclidean distance between CH and a node is relying on their situations. Consider X and Y are two nodes, i and j demonstrates two node locations. Euclidean distance is calculated based on Eq. (12):

$$d(X,Y) = \sqrt{(X_i - Y_i\ )^2 + (X_j - Y_j\ )^2} \qquad (12)$$

Each element in the distance matrix represents the difference between the CH and the node. After cluster formation, each CH is accountable for gathering the data from all the nodes in the cluster.

When a framework (of data) from all the nodes in the cluster is consummated and aggregation is performed, each CH dispatches the framework to the base station. The proceeds of reclustering and data transportation is continued for R rounds until all the nodes being dead. If the number of nodes in the cluster gets smaller than the predefined threshold, the cluster is merged with the neighboring clusters.

4.3. Tree formation of CHs and Data transmission (Phase 3)

After cluster formation, the CH sends message to all non cluster nodes in WSN which includes the CH ID, location, cluster size (for example the number of nodes in each cluster), and remaining energy. CHs also send their data and location to the base station. Base station prepares a minimum spanning tree based on the position of CH nodes so the minimum spanning tree is between CH nodes and the base station. In this plan, CHs use free area channel model to send data to the base station. In each round, the minimum distance from a vertex to another vertex is chosen based on the location of CH nodes in the tree. Combining data from several sensors used for removing the redundant transmission. Non CH nodes send their data by the framework to the CH while they are in transmission mode, so data transmission is broken into frameworks. Nodes could dispatch their data without any collision in the network. In this research, we assumed that nodes are all the time synchronized by having the base station sent out synchronization pulses to each node. When the CH receives the data from all the non CH nodes, it performs data aggregation to produce a useful data message for sending to the base station. After aggregating data, CHs transmit their resultant data along the tree (by the minimum spanning tree between CH nodes). Finally, the base station receives the final resultant data. Non CH nodes could leave clusters when its energy is finished. If any non CH node leaves, the related cluster releases it. If CH node is dead or a new node is joined to the network, the CH selection algorithm should be re-run.

In this paper, we consider two versions of DFC, DFC-1 and DFC-2. In DFC-1, a node consumes its finite energy budget during the algorithm. A specific threshold is considered in DFC-2 for the CHs. When the amount of energy of a CH passes from the specified threshold, a new CH is selected. In DFC-2, our threshold is achieved when the amount of energy of CH is reduced by half.

The flow chart of DFC algorithm is shown in Figure 2. The proposed algorithm employs the concept of cluster based and tree based data aggregation. Cluster based data aggregation is placed on the top of the flowchart and tree based data aggregation is placed in the following.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper, we proposed a hybrid protocol which is inspired from Cluster Based data aggregation and Tree Based data Aggregation. The Cluster Based method decreases energy dissipation and encounter in a local cluster. It is serious to determine the numbers of CHs that are in the WSN for maximizing the performance of energy.

In our algorithm, the number of nodes is set to 100. The sink is situated far away from the area. In Cluster based approach, we consider ten CHs (K=10) in the network. The number of considered CHs are 5, 8, 10 and 15 and R is 140. Figure 3 show that K CHs are the most optimal conditions in comparison with another CHs. The selected optimal CHs have the lowest wasted energy and dead nodes, these CHs can keep more energy. For selecting the best CHs, we have used Pareto optimal solution. Pareto optimal CHs are considered three criteria containing the remaining energy of the node, the minimum distance from the sink, and the number of adjacent nodes. The criteria are normalized in range [0, 1].

We specify the fuzzy best solutions and fuzzy worst solutions. According to these quantities, we calculate a separation rate and rating indices for the selecting node. The lifetime of the network is extended in the period of the number of cycles until the first node in the network runs out of its complete energy. CHs are chosen for each node till all the nodes expand their whole energy. In a Tree based data aggregation, an aggregated tree is constructed based on a minimum spanning tree which source nodes are thought out as leaves, so data are forwarded by the parent node for each node. The Tree based procedure has a low distance between each node and its parents, thereby wasted energy is diminished. Nevertheless, the depth of the tree is high. This hybrid method uses the advantages of the clustering and the tree structures while minimizing the disadvantages of them. Comparison of our proposed method with LEACH, CTDA and CTPEDCA is represented that the present protocol is more effective than other mentioned methods in WSNs. We use a uniform simulation environment to facilitate comparison of energy savings and consume energy between protocols. Hundred sensor nodes are randomly spread in an area and the base station is placed far away from the area. In Table 2, the parameters of our simulation are listed.

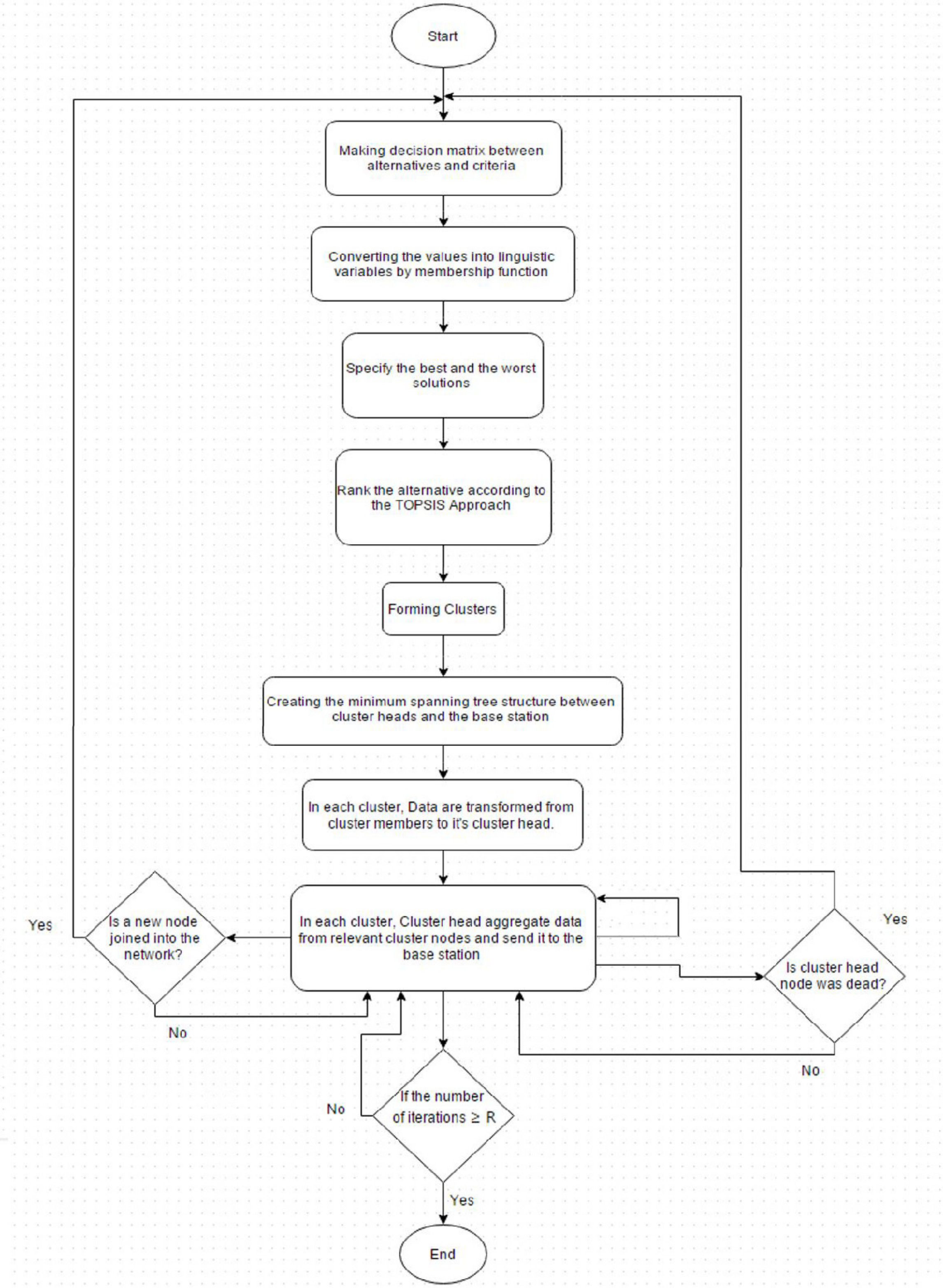

Start
Making decision matrix between alternatives and criteria
Converting the values into linguistic variables by membership function
Specify the best and the worst solutions
Rank the alternative according to the TOPSIS Approach
Forming Clusters
Creating the minimum spanning tree structure between cluster heads and the base station
In each cluster, Data are transformed from cluster members to it's cluster head.
In each cluster, Cluster head aggregate data from relevant cluster nodes and send it to the base station
Is a new node joined into the network?
Yes
No
Is cluster head node was dead?
Yes
No
If the number of iterations ≥ R
No
Yes
End

Figure 2. Flowchart of the DFC algorithm

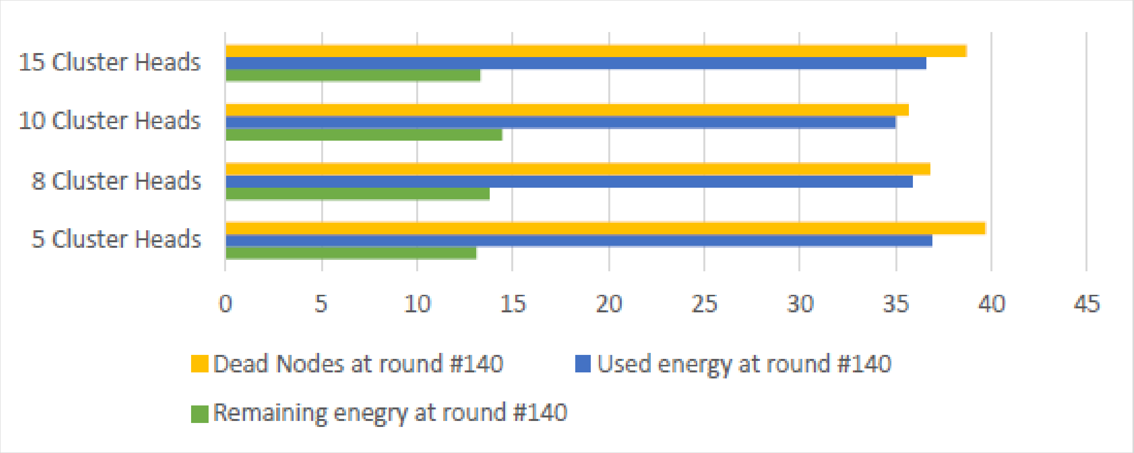


Figure 3. The effect of number of clusters on DFC based on the number of dead nodes and used energy and remaining energy at round 140.

Table 2. Simulation parameters used for WSNs

Parameter	Value
Number of nodes in the system	100
E_{elec}	50nJ / bit
ε_{fs}	10 pJ/bit/m^2
ε_{mp}	0.0013 pJ/bit/m^4
BS location	(50,200)
EDA (data aggregation)	5nJ / bit / signal
Control Packet size	800
Data Packet size	4000
R	140
K	10

A node is considered "dead" when it spent all its energy in the transferring process and also not able to send and receive the data. The simulation results of dead nodes are shown in Figure 4.

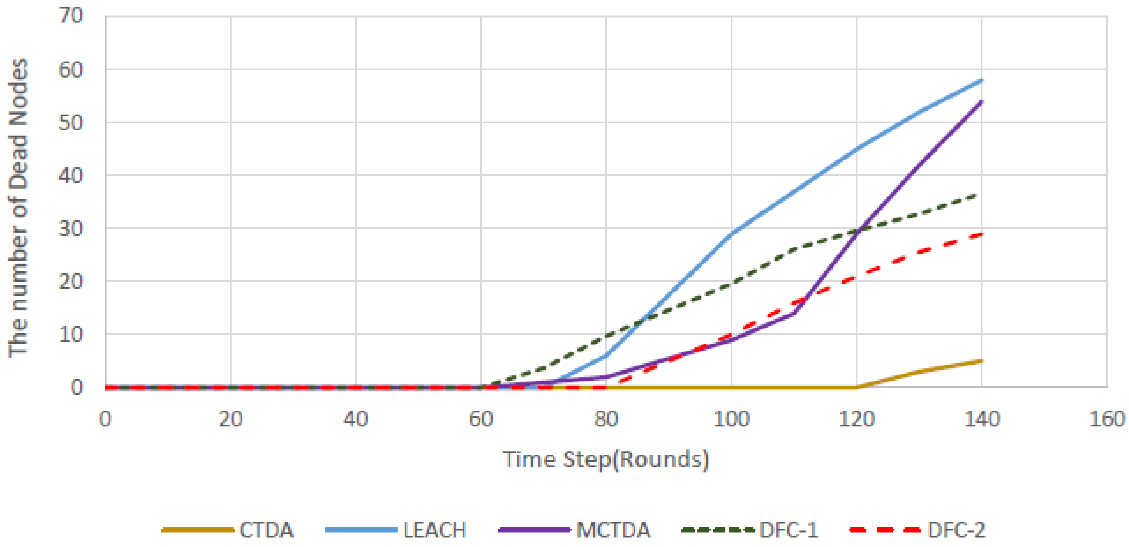


Figure 4. The number of dead nodes during the simulation.

Although the number of dead nodes in CTDA is low, but it has many disadvantages. CTDA selects CHs randomly and it does not have any calculations to select the CHs. CTDA may select a CH with very low energy or choose a CH with the least number of neighbors. The number of dead nodes in DFC-1 and DFC-2 is less than LEACH and MCTDA. This pros is because of the CHs are calculated and elected based on three criteria the remaining energy, distance to the base station and the number of neighbors around.

According to the short distance between nodes in the proposed approach, network lifetime is increased. Furthermore to decrease node solubility, DFC-1 and DFC-2 algorithms are more energy efficient all over the simulation. In DFC-2, we define a threshold for the amount of energy in CH, when the node's energy is less than the threshold, the new CH is replaced. The simulation results of residual energy are illustrated in Figure 5.

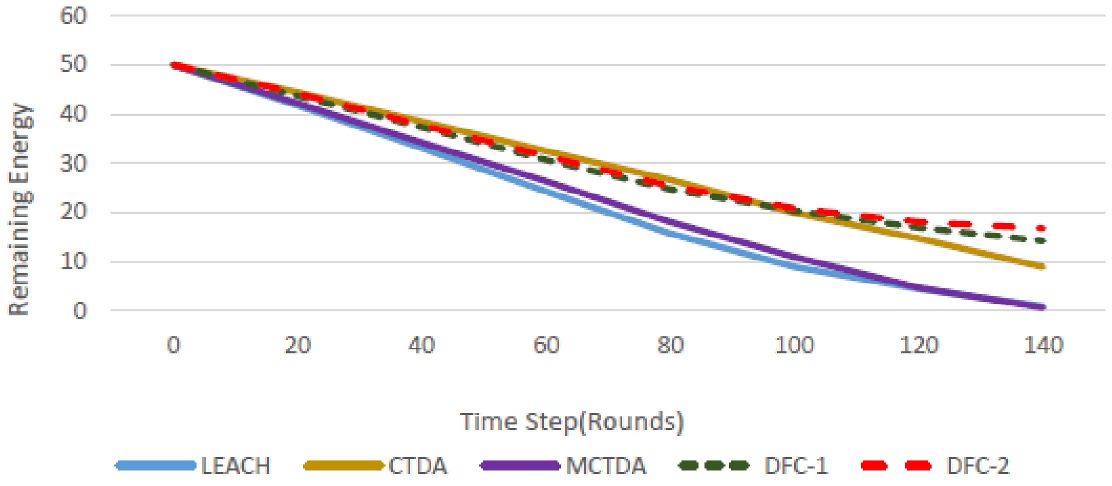


Figure 5. Remaining Energy of nodes

The results show that the remaining energy is increased. Choosing the correct CH in the proposed method make shorter distance between nodes. Nodes are selected as CHs which have the largest number of neighbors. Thus, less energy are wasted so each node can hold more energy. Energy consumption of the nodes is reduced.

6. CONCLUSIONS

Finite energy and redundant data in WSNs need data aggregation to reduce the excess number of sensors that transmit data to the base station. In this paper, we offer two main approaches in this context, included cluster based and tree based data aggregation. The fuzzy TOPSIS method is used for finding the best CHs in WSNs. Three criteria contain remaining energy, distance of the nodes from the base station and the number of neighbor nodes. These criteria are considered in order to optimize the number of CHs.

The tree based method constructs a minimum spanning tree distance between CHs and the base station, which lead to decreasing energy dissipation. We proposed an energy effective algorithm in this paper (DFC). DFC is a cluster and tree based data aggregation and is compared with LEACH, CTDA and MCTDA protocols. The conclusions of this simulation demonstrate that DFC considerably saves energy of nodes which increases the network lifetime compared to the other protocols.

REFERENCES

[1] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: a survey. Computer networks, 38(4), 393-422.

[2] Yu, J. Y., & Chong, P. H. J. (2005). A survey of clustering schemes for mobile ad hoc networks. Communications Surveys & Tutorials, IEEE, 7(1), 32-48.

[3] Abbasi, A. A., & Younis, M. (2007). A survey on clustering algorithms for wireless sensor networks. Computer communications, 30(14), 2826-2841.

[4] Asemani, M., & Esnaashari, M. (2015). Learning automata based energy efficient data aggregation in wireless sensor networks. Wireless Networks, 1-19.

[5] Selvin, S. V., & Kumar, S. M. (2012). Tree based energy efficient and high accuracy data aggregation for wireless sensor networks. Procedia Engineering, 38, 3833-3839.

[6] Azad, P., & Sharma, V. (2013). Cluster head selection in wireless sensor networks under fuzzy environment. ISRN Sensor Networks, 2013.

[7] Baykasoğlu, A., & Gölcük, İ. (2015). Development of a novel multiple-attribute decision making model via fuzzy cognitive maps and hierarchical fuzzy TOPSIS. Information Sciences, 301, 75-98.

[8] Rathod, M. K., & Kanzaria, H. V. (2011). A methodological concept for phase change material selection based on multiple criteria decision analysis with and without fuzzy environment. Materials & Design, 32(6), 3578-3585.

[9] Yang, T., & Hung, C. C. (2007). Multiple-attribute decision making methods for plant layout design problem. Robotics and computer-integrated manufacturing, 23(1), 126-137.

[10] Lee, M., & Wong, V. W. (2005). An energy-aware spanning tree algorithm for data aggregation in wireless sensor networks, IEEE Pacific Rim Conference on Computers and signal Processing and Communications, 300-303

[11] Rajagopalan, R., & Varshney, P. K. (2006). Data aggregation techniques in sensor networks: A survey.

[12] Abdullah, L., & Najib, L. (2014). A new type-2 fuzzy set of linguistic variables for the fuzzy analytic hierarchy process. Expert Systems with Applications, 41(7), 3297-3305.

[13] Sun, C. C. (2010). A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods. Expert systems with applications, 37(12), 7745-7754.

[14] Akkari, W., Bouhdid, B., & Belghith, A. (2015). LEATCH: Low Energy Adaptive Tier Clustering Hierarchy. Procedia Computer Science, 52, 365-372.

[15] Sajedi, H., & Saadati, Z. (2014). A Hybrid Structure for Data Aggregation in Wireless Sensor Network. Journal of Computational Engineering, 2014.

[16] Ranjani, S. S., Krishnan, S. R., & Thangaraj, C. (2012, April). Energy-efficient cluster based data aggregation for wireless sensor networks. International Conference on Recent Advances in Computing and Software Systems, 174-179.

[17] Wang, W., Wang, B., Liu, Z., Guo, L., & Xiong, W. (2011). A cluster-based and tree-based power efficient data collection and aggregation protocol for wireless sensor networks. Information technology journal, 10(3), 557-564.

[18] Richard, W. G. (2009). Extending LEACH routing algorithm for Wireless Sensor Network. Data Communications Engineering, Makerere University.

[19] Batra, N., Jain, A., & Dhiman, S. (2011). An optimized energy efficient routing algorithm for wireless sensor network. International Journal of Innovative Technology and Creative Engineering.

[20] Jorio, A., El Fkihi, S., Elbhiri, B., & Aboutajdine, D. (2015). An Energy-Efficient Clustering Routing Algorithm Based on Geographic Position and Residual Energy for Wireless Sensor Network. Journal of Computer Networks and Communications.

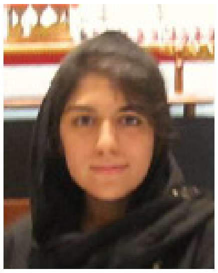
[21] Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000). Energy-efficient communication protocol for wireless microsensor networks, annual Hawaii international conference on System sciences.

[22] Chauhan, A., & Vaish, R. (2013). Pareto optimal microwave dielectric materials. Advanced Science, Engineering and Medicine, 5(2), 149-155.

[23] Kasprzak, E. M., & Lewis, K. E. (2001). Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. Structural and Multidisciplinary Optimization, 22(3), 208-218.

[24] Zanakis, S. H., Solomon, A., Wishart, N., & Dublish, S. (1998). Multi-attribute decision making: A simulation comparison of select methods. European journal of operational research, 107(3), 507-529.

AUTHORS



Arezoo Abasi was born in Tehran, Iran in 1994. She studies Computer Science at University of Tehran since 2012. She was awarded Khaje Nasir price in 2010. Her interests include artificial intelligence, soft computing and wireless sensor network.

Hedieh Sajedi received a B.Sc. degree in Computer Engineering from AmirKabir University of Technology in 2003, and M.Sc. and Ph.D degrees in Computer Engineering (Artificial Intelligence) from Sharif University of Technology, Tehran, Iran in 2006 and 2010, respectively. She is currently an Assistant Professor at the Department of Computer Science, University of Tehran, Iran. Her research interests include Computer Networks, Machine Learning, and Signal Processing.

DIAGNOSIS OF RHEUMATOID ARTHRITIS USING AN ENSEMBLE LEARNING APPROACH

Zahra Shiezadeh[1], Hedieh Sajedi[2] and Elham Aflakie[3]

[1]Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, Bushehr, Iran
zshiezadeh@gmail.com
[2]Mathematics, Statistics and Computer Science School,
College of Science, University of Tehran, Tehran, Iran
hhsajedi@ut.ac.ir
[3]Rheumatology Research Center,
Shiraz University of Medical Sciences, Shiraz, Iran
aflakye@sums.ac.ir

ABSTRACT

Rheumatoid arthritis is one of the diseases that its cause is unknown yet; exploring the field of medical data mining can be helpful in early diagnosis and treatment of the disease. In this study, a predictive model is suggested that diagnoses rheumatoid arthritis. The rheumatoid arthritis dataset was collected from 2,564 patients referred to rheumatology clinic. For each patient a record consists of several clinical and demographic features is saved. After data analysis and pre-processing operations, three different methods are combined to choose proper features among all the features. Various data classification algorithms were applied on these features. Among these algorithms Adaboost had the highest precision. In this paper, we proposed a new classification algorithm entitled CS-Boost that employs Cuckoo search algorithm for optimizing the performance of Adaboost algorithm. Experimental results show that the CS-Boost algorithm enhance the accuracy of Adaboost in predicting of Rheumatoid Arthritis.

KEYWORDS

Data Mining, Adaboost, Cuckoo's Algorithm, Predictive Model, Rheumatoid Arthritis, Decision Tree.

1. INTRODUCTION

Rheumatoid arthritis (RA), is one of the arthritis that causes inflammation, pain and swelling in the joints. Usually it is chronic and can cause long term damage or deformation of the joints. One out of every hundred people is in some way affected by RA in the life [1, 2]. The cause of RA is still unknown. If RA is diagnosed in its early stages, it can be easily controlled. Usually

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 139–148, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51512

RA is diagnosed when severe symptoms appear in the patient and disease need more aggressive treatment [3].

A variety of methods has been developed for the early diagnosis of RA [3, 4], such as 2010 ACR/EULAR classification criteria [5] and the van der Helm–van Mil (vHvM) score [6]. Currently, clinical experience is the basis of the diagnosis of RA using certain RA disease classification criteria. Precise and accurate assessment of RA symptoms can avoid permanent damage to the patient's joints and bones, and also have influence on patients' quality of life. Rheumatoid arthritis is an area of medicine that has been less considered from the perspective of data mining.

In recent years, data mining using electronic medical records has been very popular and is expected to improve the accuracy of diagnosis and quality treatment using data mining techniques. Developing a model that can infer predicted class is the purpose of the prediction model [7]. In this study, a predictive model for automatic detection of rheumatoid arthritis was developed using an integrated approach. The study relates to the factors that predict disease, data mining, classification techniques, and a database was created for patients in the rheumatology clinic of Shiraz university of medical sciences.

2. RELATED WORKS

In most of the data mining studies that were investigated, more attention has been paid to several medical fields, including RA [8, 9, 10], cardiovascular diseases [11, 12, 13, 14, 15, 16], cancer [17, 18, 19, 20], lung [21, 22, 23], traumatic brain injury [24, 25, 26] and diabetes [27, 28, 29].

In our study, we investigate the data mining applications in the prediction of RA diseases, and associated classification techniques. Previous researches have studied various attributes and different method to establish disease prediction models. For instance, Pinar Yildirim et al.[8] use textmining to discover similar attribute among RA patients whereas Cader et al. [30] used the 2010 and 1987 ACR/EULAR criteria for the prediction of RA patients. Huizinga et al. [31] examined nine clinical variables to construct prediction rules.

3. THE PROPOSED ALGORITHM

Adaboost algorithm was the first practical boosting algorithm and remains one of the most widely used and studied, and it is applied in numerous fields. This ensemble method is selected to apply to the database and deriving results from it [32, 33, 34, 35]. The next step is to improve the accuracy of modeling, so the Adaboost algorithm [36, 37] is combined with Cuckoo Search algorithm [38, 39] and CS-Boost algorithm is proposed. Cuckoo search (CS) is an optimization algorithm developed by Xin-she Yang and Suash Deb in 2009. In the proposed algorithm, the Decision Stump has been used as weak learners. The tree is designed as a weak learner or a base learner for bagging or boosting techniques [38], and it makes one level decision tree for classical or numerical data sets.

4. RESEARCH METHODS AND MATERIALS

Considering advantages of the CRISP-DM model for knowledge discovery, in this study the various stages of the model were implemented. The CRISP-DM model is explained as a procedure of the cross industry standard for developing data mining projects. This is one of the most widely used data mining techniques to discover knowledge, and the six phases implement the model are: The business understanding, data understanding, data preparation, and modeling, evaluation, and deployment phases. A detail of this methodology is available in [27]. In the following, we describe our implementation of each phase.

4.1. Business Understanding:

This study provides a model to predict rheumatoid arthritis among patients referred to the Rheumatology clinic. Due to the nature of rheumatic diseases, there is a wide range of overlapping symptoms in RA. A predictive model can be effective in medical knowledge in this field and promote the health of the society [41].

4.2. Data Understanding:

The cohort consisted of patients referred to the rheumatology clinic at Shiraz University of Medical Sciences during the study period. We identified 2564 patients who were admitted to the clinic with arthritis diagnoses. We constructed a new data set for the arthritis and for each patient, we have saved records consist of the demographic and clinical data. The final data set has contained more than 600 attributes such as demographic data, lab data, treatments, and physical exams, symptoms, past history and having pain, redness, and tenderness and… in the patient's joints. Finally, we categorized data values and derived new fields from existing data. These features were changed to categorical attributes for better analysis and to obtain good results. More than 72 features were selected due to preliminary feature selection and physicians' opinions. Table1 shows these features, their values and data types.

4.3. Data Preparation:

In the data preparation phase, the data were preprocessed. The preprocessing phase includes the following steps:

- *Data Cleaning*

- *Constructing New Data*: New fields are derived from existing ones. The total number of joints, MCP count pj15 to pj24, PIP count pj25 to pj40, DIP count pj41 to pj56, MTP count pj61 to pj75 and BMI that is calculated from weight and height. Figure1 shows the joint and their locations.

Table 1: Data set features and their values and data types.

Attributes	Values	Data type	Attributes	Values	Data type
Code		Integer	**Height**		Integer
Sex	{Male, Female}	Binominal	**Weight**		Integer
Age		Integer	**Disease duration**		Integer
Birth place		Nominal	**pj2[1]-pj32**	{Yes, No}	Binominal
Marital status	{Single, Married, Divorced, Widowed}	Polynomial	**Pj37-pj49**	{Yes, No}	Binominal
Education	{Diploma, High School, Associate, BS/BA, None, Intermediate, Primary School, MS/BA, Doctoral}	Polynomial	**Pj57-pj75**	{Yes, No}	Binominal
Job	{Disabled, Full Time Employed, Housewife, Part-time Employed, Retired, Student, Unemployed}	Polynomial	**ESR**		Integer

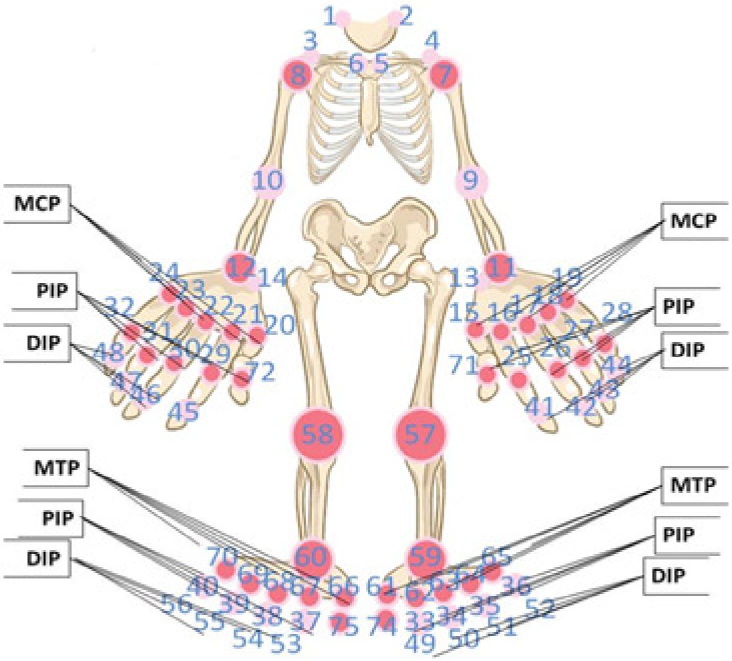


Figure 1: The joints and their location in the patient's body

[1] Patient Joint

- *Feature Selection:* At this stage, different methods for selecting effective features, in three steps have been taken. At this stage only features influencing the target field are selected as the input for modeling phase. 1) In this way, using feature selection techniques such as Chi Squared, CFS, Gain Ratio, Info Gain, OneR and Relief, the main features from each technique separately extracted [42, 43]. To increase the accuracy of assessing, 10-fold Cross Validation is used. 2) The Feature Selection nodes in SPSS Modeler software are used to select the most important features. In this way, by eliminating features that have small variances, ranked features and 11 features are selected as a subset of them. 3) The features are presented to specialists and they ranked them. The results from these three steps are integrated and finally 18 features are entered to modeling phase. Table2 shows the feature selection results.

Table 2: Features for modeling phase

Feature	Rank	Feature	Rank
Joint Count	1	MCP	10
ESR	2	pj8	11
PIP	3	pj10	12
pj9	4	pj12	13
pj11	5	Age	14
DIS	6	Duration	15
pj58	7	MTP	16
Sex	8	pj59	17
pj57	9	Marital Stat	18

4.4. Modeling:

The learning model in this study is supervised method, considering the goal field which is diagnosed by specialists as well as finding the most important factors influencing the diagnosis of rheumatoid arthritis. In fact, goal feature has two distinct values, susceptibility to rheumatoid arthritis (RA) and other rheumatic diseases (Other), so the nature of data mining tends to classification. Therefor applying classification algorithms that extract the rules and determine the relationship between individual features and goal feature is the main parts of the model. In this study, first C4.5, CHAID, ID3, W-J48 and Adaboost algorithms are implemented on the dataset. Then SVM, KNN and Adaboost algorithm with Decision Stump as a weak learner are implemented using MATLAB software on the dataset. In the implementation of these algorithms, the doctor's diagnosis was goal feature and other features that are selected in the selection phase are considered as the input features. By implementing the above steps, the ensemble algorithm, Adaboost accuracy in modeling was higher than other methods using the combination of weak classifier.

4.5. Evaluation:

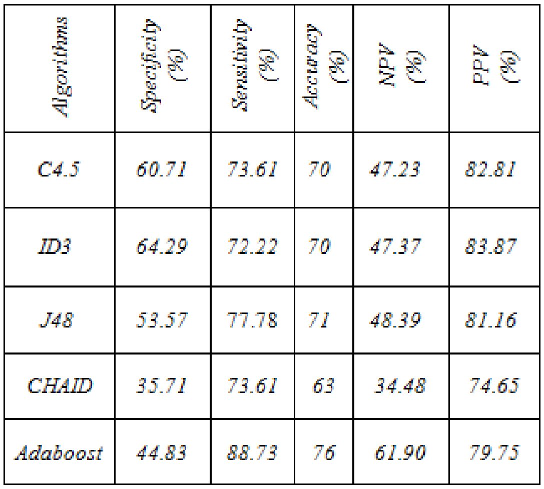
The algorithms are applied to the data set using stratified 5-fold validation in order to assess the performance of classification techniques for predicting a class. Evaluation criteria in classification problems are accuracy, sensitivity, specificity, PPV and NPV that are achieved using confusion matrix.

Table 3 : A Confusion matrix Table

Confusion Matrix	Other	RA	Class precision
Other	TN	FN	NPV=TN/(TN+FN)
RA	FP	TP	PPV=TP/(TP+FP)
Class recall	Specificity=TN/(TN+FP)	Sensitivity=TP/(TP+FN)	
Accuracy	(TP+TN)/(TP+TN+FP+FN)		

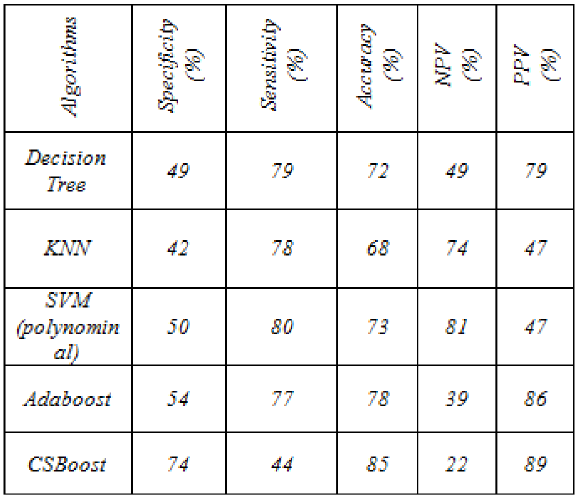
PPV: It denotes the percentage of RA predictions that are correct. Recall / Sensitivity: It denotes the percentage of RA labeled instances that were predicted as RA. Specificity: It denotes the percentage of Other labeled instances that were predicted as Other. Accuracy: It denotes the percentage of predictions that are correct. Table3 shows the comparison of decision tree that the Adaboost with J48 algorithm as the base learner has maximum accuracy and sensitivity. Table4 shows the comparison of classification algorithms. The Adaboost algorithm implemented with decision stamp as base learner.The proposed CS-Boost has maximum accuracy and the minimum sensitivity.

Table 3: Decision Tree Comparison



Algorithms	*Specificity (%)*	*Sensitivity (%)*	*Accuracy (%)*	*NPV (%)*	*PPV (%)*
C4.5	*60.71*	*73.61*	*70*	*47.23*	*82.81*
ID3	*64.29*	*72.22*	*70*	*47.37*	*83.87*
J48	*53.57*	*77.78*	*71*	*48.39*	*81.16*
CHAID	*35.71*	*73.61*	*63*	*34.48*	*74.65*
Adaboost	*44.83*	*88.73*	*76*	*61.90*	*79.75*

Table 4: Classification model comparison



Algorithms	*Specificity (%)*	*Sensitivity (%)*	*Accuracy (%)*	*NPV (%)*	*PPV (%)*
Decision Tree	*49*	*79*	*72*	*49*	*79*
KNN	*42*	*78*	*68*	*74*	*47*
SVM (polynomial)	*50*	*80*	*73*	*81*	*47*
Adaboost	*54*	*77*	*78*	*39*	*86*
CSBoost	*74*	*44*	*85*	*22*	*89*

Table 3 shows that the AdaBoost algorithm with J48 algorithm as the weak learner has the highest Sensitivity 88.73 percent and can diagnose RA correctly for a rheumatic patient. The ID3 algorithm on this dataset for patients with other arthritis diseases will have discretion 64.29 percent, the highest Specificity compared to other algorithms. If you have RA patients, the ID3 algorithm detects the correct model, 83.87 percent and has the highest percision. If patients have diseases other than RA, AdBoost 61.90 percent can recognize other diseases correctly that is the highest percentage among the other algorithms. Among these algorithms, AdaBoost algorithm has the highest accuracy 76 percent.

Table 4 shows that the CSBoost algorithm has the highest PPV, 89 percent. The SVM algorithm on this dataset has the highest Specificity, 80 percent compared to other algorithms. If patients have diseases other than RA, SVM 81 percent can recognize other diseases correctly that is the highest percentage among the other algorithms. Among these algorithms, CSBoost algorithm has the highest accuracy 85 percent.

5. DISCUSSION

This study, carried out along with the 2564 patients refer to the Rheumatology clinic in Shiraz university of medical sciences, and used data mining technology to construct a rheumatoid arthritis disease predictive model. A total of 300 valid sample patients was acquired from this database, the data on the patients were collected for classification study, which included their physical exam results, symptoms, lab data results, patient history, demographic data and diagnoses. Data mining technologies adopted in this study were decision tree, c4.5, Id3, Chaid, WJ48, SVM, KNN and boosting algorithm (Adaboost). In comparison, of data mining technology, this study used sensitivity and accuracy indicators to evaluate classification efficiency of different algorithms. After comparing classification accuracy, Adaboost was the best classification algorithm in this study.

The optimum RA disease predictive model obtained in this study adopts CS-Boost as classification algorithm, 18 attributes as attribute input mode, and its classification efficiency: sensitivity indicator = 44% and accuracy indicator = 85%. 18 major influence factors were recognized for accurately predicting RA disease but education, BMI, occupation and birthplace were less important as other factors that is similar to the 2010 ACR/EULAR Classification Criteria. The research results could not be comparable with the other similar mining researches in RA such as [8, 9, 10] because they have used text mining and their search result was related to RA but were different from this study. In addition, 20 diagnosis, classification rules were extracted from this predictive model, and confirmed by three RA specialists to be conformable with the current clinical medical condition and have reference value in diagnosis and prediction of RA disease.

This study has some weak points. The research carried on patient that the treatment was started for their disease. It is suggested that in the future work the dataset can be collected from new case patients. For this research, there are 2564 records patients, but 357 of these records was suitable for this study so in the future work classification can be applied in more patients' records to get more precise and accurate results.

6. CONCLUSION

Rheumatoid arthritis, is one of the diseases that its cause is not known yet, data mining can help the medical field in order to provide early diagnosis and treatment of this disease. The aim of this study is to provide predictive models for the diagnosis of rheumatoid arthritis. The data were collected from patients referred to the rheumatology clinic of Shiraz University of Medical Sciences. Next the data is preprocessed. Decision tree algorithms for the modeling are applied such as C4.5, ID3, CHAID, J48, SVM and Adaboost. Then the Adaboost algorithm is combined with a Cuckoo Search algorithm and CSBoost algorithm is proposed. The optimum RA disease predictive model obtained in this study adopts CSBoost as classification algorithm. Comparison of the models has shown that CSBoost has the highest accuracy among them. The results indicate that elbow and knee joints, gender, number of joints and ESR test result have the most impact in the diagnosis of rheumatoid arthritis. The models can be applied in a computer software to predict rheumatoid arthritis and become a decision support for physicians.

REFERENCES

[1] Scott DL, Wolfe F, Huizinga TW,"Rheumatoid arthritis", Lancet, Volume 376 (9746),pp.1094–108, 2010, Sep 25.

[2] Majithia V, Geraci SA. "Rheumatoid arthritis: diagnosis and management". American Journal of Medicine. Volume 120- No11, pp. 936–9, 2007.

[3] Chin CY, Weng MY, Lin TC, Cheng SY, Yang YHK, Tseng VS. "Mining Disease Risk Patterns from Nationwide, Clinical Databases for the Assessment of Early Rheumatoid Arthritis Risk". PLoS ONE. Volume 10- N0-4, 2015.

[4] RACGP. "Clinical guideline for the diagnosis and management of early rheumatoid arthritis". The Royal Australian College of General Practitioners, 1 Palmerston Crescent, South Melbourne, Vic 3205 Australia, 2009

[5] Cader MZ, Filer A, Hazlehurst J, de Pablo P, Buckley CD, Raza K. "Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort. Annals of the Rheumatic Diseases". Volume 70- No. 6. pp. 949–55, 2011.

[6] van der Helm-vanMil AHM, le Cessie S, van Dongen H, Breedveld FC, Toes REM, Huizinga TWJ. "A prediction rule for disease outcome in patients with Recent-onset undifferentiated arthritis: How to guide individual treatment decisions". Arthritis & Rheumatism.; Volume 56- No.2, pp. 433–40. 2007 Feb.

[7] A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.

[8] Bedran Z, Quiroz C, Rosa J, Catoggio LJ, Soriano ER. "Validation of a Prediction Rule for the Diagnosis of Rheumatoid Arthritis in Patients with Recent Onset Undifferentiated Arthritis". International Journal of Rheumatology. 548502, 2013.

[9] G. Zheng, M. Jiang, C. Lu, H. Guo, J. Zhan و A. Lu, "Exploring the biological basis of deficiency pattern in rheumatoid arthritis through text mining," BIBM Workshops;, 2011.

[10] P. Yildirim, Ç. Çeken, R. Hassanpour , M. R. Tolun, "Prediction of Similarities Among Rheumatic Diseases," J. Medical Systems, Vol.36., No 3, pp. 1485-1490, 2012

[11] Jyoti Soni, Ujma Ansari, Dipesh Sharma," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, pp. 43-48, March 2011.

[12] Chaitrali S. Danger, Sulabha S. Apte, —"Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[13] Sellappan Palaniappan, Rafiah Awang," Intelligent Heart Disease Prediction System Using Data Mining Technique", 978-1-4244-1968-5/08 IEEE, 2008

[14] Jyoti Soni, Sunita Soni et al., —Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction‖; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[15] Vanisree K, Jyothi Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications (0975 – 8887) Volume 19– No.6, April 2011

[16] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, Volume.31 No.4, pp. 642-656,2009

[17] Ng T, Chew L, Yap CW. A Clinical Decision Support Tool To Predict Survival in Cancer Patients beyond the 120 Days after Palliative Chemotherapy. Journal of Palliative Medicine; Volume 15- No 8, pp. 863–9. 2012 Aug.

[18] Shweta Kharya," USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, 2012 April

[19] Lee SM, Kang JO, Suh YM. ,"Comparison of hospital charge prediction models of colorectal cancer patients: neural network vs. decision tree models";J Korean MED Sci., Volume 19 (5), pp. 677-81, 2004 Oct

[20] Chien CW, Lee YC, Ma T, Lee TS, Lin YC, Wang W. and Lee WJ," The application of artificial neural networks and decision tree model in predicting post-operative complication for gastric cancer patients". Hepatogastroenterology, Volume 55, pp. 1140- 1145, 2008.

[21] Wytske A. Altenburg, Mathieu H.G. de Greef, Nick H.T. Ten Hacken, Johan B. Wempe," A better response in exercise capacity after pulmonary rehabilitation in more severe COPD patients", Respiratory Medicine volume 106, pp. 694-700,2012.

[22] Behnke M, Wewel AR, Kirsten D, Jorres RA, Magnussen H.,"Exercise training raises daily activity stronger than predicted from exercise capacity in patients with COPD", Respir Med Jun;Volume 99(6):pp.711-717, 2005.

[23] Garrod R, Marshall J, Barley E, D. Jones PW.," Predictors of success and failure in pulmonary rehabilitation". EUR Respir J, Volume 27 (4), pp. 788-94,2006 April.

[24] A. Marcano-Cedeño , Paloma Chausa a, Alejandro García, César Cáceres, Josep M. Tormos, Enrique J. Gómez," Data mining applied to the cognitive rehabilitation of patients with acquired brain injury", Expert Systems with Applications, Volume 40, pp. 1054–1060,2013.

[25] Pang, B. C., Kuralmani, V., Joshi, R., Hongli, Y., Lee, K. K., Ang, B. T., et al,"Hybrid outcome prediction model for severe traumatic brain injury", Journal of Neurotrauma,Volume 24(1),pp. 136–146,2007.

[26] Rughani, A. I., Dumont, T. M., Lu, Z., Josh Bongar, M. S., Horgan, M. A., Penar, P. L.,et al," Use of an artificial neural network to predict head injury outcome." ,Journal of Neurosurgery, Volume 113 (3), pp. 585–590,2011

[27] S.Priya, R.R.Rajalaxmi," An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network", ICON3C,pp.26-29,2012

[28] B.M. Patil, R.C. Joshi, Durga Toshniwal, "Hybrid Prediction Model for Type-2 Diabetic patients", Expert Systems with Applications, Science direct, pp. 8102-8108,2010.

[29] Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT."Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses". PLoS ONE 9(6): e98587. doi:10.1371/journal.pone.0098587,2014.

[30] Cader MZ, Filer A, Hazlehurst J, de Pablo P, Buckley CD, Raza K. Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort. Annals of the Rheumatic Diseases. 2011; 70(6):949–55.

[31] Huizinga TWJ, van der Helmvan Mil AHM. Prediction and prevention of rheumatoid arthritis. Revista Colombiana de Reumatología. 2007; 14:106–14

[32] Freund, Y., "An Adaptive Version of the Boost by Majority Algorithm," Machine Learning, 43, 293-318, 2001.

[33] Freund, Y., and Schapire, R., "A Short Introduction to Boosting," Journal of the Japanese Society for Artificial Intelligence, 14, 771-780, 1999.

[34] Freund, Y., and Schapire, R., "A Decision-Theoretic Generalization of On-LineLearning and an Application toBoosting," Journal of Computer and System Sciences, 55(1), 119–139, 1997.

[35] Friedman, J. H., "Recent Advances in Predictive (Machine) Learning," Journal of Classification, 23, 175-197, 2006.

[36] E. AhmedSharaf , M. A. Moustafa, M. Harb و A. Emara, "ADABOOST ENSEMBLE WITH SIMPLE GENETIC ALGORITHM FOR STUDENT PREDICTION MODEL," International Journal of Computer Science & Information Technology (IJCSIT), Volume 5(2), pp. 73-85, 2013.

[37] M. Billah, Symptom Analysis of Parkinson Disease using SVM-SMO and Ada-Boost Classifiers, BRAC University, Dhaka, Bangladesh, 2014.

[38] X.-S. Yang, S. Deb, "Cuckoo Search via L´evy Flights," IEEE World Congress on Nature & Biologically Inspired Computing;, pp. 210-215, 2009.

[39] R. Rajabioun, "Cuckoo Optimization Algorithm," Applied Soft Computing Journal;, Vol.11, pp. 5508, 2011.

[40] L. Reyzin, R. E. and Schapire, “How Boosting the Margin Can Also Boost Classifier Complexity;,” Proceedings of the 23rd international conference on Machine Learning;, pp. 753-760;, 2006.

[41] Linda Miner et al., “Practical Predictive Analytics and Decisioning Systems for Medicine”, ISBN: 978-0-12-411643-6, 2015.

[42] P. Saengsiri,S.N.Wichian,P.Meesad,andU.Herwig,“Comparison of hybrid feature selection models on gene expression data,” in 8th International Conference on ICT and Knowledge Engineering, pp.13-18, 2010.

[43] Shoushan Li et al , “A frame work of feature Selection Methods for Text categorization” , Proceedings of 47th Annual meeting of ACL & 4th ICCNLP of AFNLP , pp 692-700, 2009.

LEARNING SCHEDULER PARAMETERS FOR ADAPTIVE PREEMPTION

Prakhar Ojha[1], Siddhartha R Thota[2], Vani M[1] and Mohit P Tahilianni[1]

[1]Department of Computer Engineering,
National Institute of Technology Karnataka, Surathkal, India
[2]Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, India
prakharojha992@gmail.com, sddhrthrt@gmail.com, vani@nitk.edu.in,
tahilianni@nitk.edu.in

ABSTRACT

An operating system scheduler is expected to not allow processor stay idle if there is any process ready or waiting for its execution. This problem gains more importance as the numbers of processes always outnumber the processors by large margins. It is in this regard that schedulers are provided with the ability to preempt a running process, by following any scheduling algorithm, and give us an illusion of simultaneous running of several processes. A process which is allowed to utilize CPU resources for a fixed quantum of time (termed as timeslice for preemption) and is then preempted for another waiting process. Each of these 'process preemption' leads to considerable overhead of CPU cycles which are valuable resource for runtime execution. In this work we try to utilize the historical performances of a scheduler and predict the nature of current running process, thereby trying to reduce the number of preemptions. We propose a machine-learning module to predict a better performing timeslice which is calculated based on static knowledge base and adaptive reinforcement learning based suggestive module. Results for an "adaptive timeslice parameter" for preemption show good saving on CPU cycles and efficient throughput time.

KEYWORDS

Reinforcement Learning, Online Machine Learning, Operating System, Scheduler, Preemption

1. INTRODUCTION

Scheduling in operating systems is based on time-sharing techniques where several processes are allowed to run "concurrently" so that the CPU time is roughly divided into "slices", one for each runnable process. A single core processor, which can run only one process at any given instant, needs to be time multiplexed for running more processes simultaneously. Whenever a running process is not terminated upon exhausting its quantum time slice, a switch takes place where another process in brought into CPU context. *Linux processes* have the capability of preemption [8]. If a process enters the RUNNING state, the kernel checks whether its priority is greater than the priority of the currently running process. If this condition is satisfies then the execution is interrupted and scheduler is invoked to select the process that just became runnable or any

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 149–162, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51513

another process to run. Otherwise, a process is also to be preempted when its time quantum expires. This type of time sharing relies upon the interrupts and is transparent to processes.

A natural question to ask would be - *How long should a time quantum last?* The duration, being critical for system performances, should be neither too long nor too short [8]. Excessively short periods will cause system overhead because of large number of task switches. Consider a scenario where every task switch requires 10 milliseconds and the time slice is also set to 10 milliseconds, then at least 50% of the CPU cycles are being dedicated to task switch. On the other hand if quantum duration is too long, processes no longer appear to be executed concurrently[15]. For instance, if the quantum is set to five seconds, each runnable process makes progress for about five seconds, but then it stops for a very long time (typically, five seconds times the number of runnable processes). When a process has exhausted its time quantum, it is preempted and replaced by another runnable process. Every time a process is pushed out to bring in another process for execution (referred as context switch) several other elementary operations like swap-buffers, pipelines clearances, invalidate cache etc. take place making process switch a costly operation. [16] So preemption of a process leads to considerable overhead.

As there does not exist any direct relation between timeslice and other performance metrics, our work proposes a machine-learning module to predict a better performing timeslice. The proposed adaptive time slice for preemption displays improvements in terms of the the total time taken (Turnaround Time) after the submission of process to its completion, in-return creating more processor ticks for future. Most of present work has hard-wired classifiers which are applicable only to certain types of jobs. Having a reinforcement learning agent with reward-function, which learns over time, gives the flexibility of adapting to dynamic systems. The subsequent sections will briefly discuss the fundamentals of reinforcement learning framework, which strives to continuously improve self by learning in any new environment.

Following sections in this paper are organized as follows: Section 2 gives an overview of the related previous works and Section 3 explains the theory of our reinforcement learning framework. Section 4 shows how we approach the problem in hand by proposing a novel design, integrate RL modules and run simmulations and then followed by implementation details of knowledge base created and self-learning systems in Section 5. The results and analysis of our system's performace is evaluated in Section 6 followed by conclusions and discussions in Section 7.

2. RELATED WORKS

Attempts have been made to use historical-data and learn the timeslice parameter, which judges the preemption time for a given process, and make it more adaptive. Below section briefly discusses earlier works in relevant fields by applying machine learning techniques to CPU resources and Operating system parameters.

To remember the previous execution behaviour of certain well-known programs, [10] studies the process times of programs in various similarity states. The knowledge of the program flow sequence (PFS), which characterizes the process execution behaviour, is used to extend the CPU time slice of a process. They also use thresholding techniques by scaling some feature to determine the time limit for context switching. Their experimental results show that overall processing time is reduced for known programs. Works related to Thread schedulers on multi core systems, using Reinforcement learning, assigns threads to different CPU cores [6], made a case that a scheduler must balance between three objectives: optimal performance, fair CPU sharing and balanced core assignment. They also showed that unbalanced core assignment results in performance jitter and inconsistent priority enforcement. A simple fix that eliminates jitter and presents a scheduling framework that balances these three objectives by algorithm based on reinforcement learning was explored.

The work in [7] has addressed similar problem based on making fixed classifiers over hand picked features. Here timeslice values were tried against several combination of attributes and patterns emerged for chosing better heuristic. However, their approach was compatible to only few common processes like random number generation, sorting etc. and unlike our work, not universally adaptive for any application. Reward based algorithms and their use in resolving the lock contention has been considered as scheduling problem in some the earlier works[2]. These hierarchal spin-locks are developed and priority assigned to processes to schedule the critical-section access.

Application run times are predicted using historical information in [1]. They derive predictions for run times of parallel applications from the run times of similar applications that have executed in the past. They use some of the following characteristics to define similarity: user, queue, load leveler script, arguments, network adapter, number of nodes, maximum run time, submission time, start time, run time. These characteristics are used to make a template which can find the similarity by matching. They use genetic algorithms (GA), which are are well known for exploring large search spaces, for identifying good templates for a particular workload.

Statistical Regression methods, which work well on numeric data but not over nominal data, are used for prediction [5]. An application signature model for predicting performance is proposed in [4] over a given grid of resources. It presents a general methodology for online scheduling of parallel jobs onto multi-processor servers, in a soft real-time environment. This model introduces the notion of application intrinsic behaviour to separate the performance effects of the runtime system from the behaviour inherent in the application itself. Reinforcement Learning is used for tuning its own value function which predicts the average future utility per time step obtained from completed jobs based on the dynamically observed state information. From this brief review of related literature, we draw the following conclusions:

- It is possible to profitably predict the scheduling behaviour of programs. Due to the varied results in all above discussed works, we believe that the success of the approach depends upon the ML technique used to train on previous programs execution behaviour.

- A suitable characterization of the program attributes (features) is necessary for these automated machine learning techniques to succeed in prediction.

In specific to using reinforcement learning in realms of scheduling algorithms, most of the work is concentrated around ordering the processes like to learn better permutations of given list of processes, unlike our work of parameter estimation.

3. Reinforcement Learning Framework

Reinforcement learning (RL) is a collection of methods for approximating optimal solutions to stochastic sequential decision problems [6]. An RL system does not require a teacher to specify correct actions, instead, it tries different actions and observes their consequences to determine which actions are best. More specifically, in any RL framework, a learning agent interacts with its environment over a series of discrete time steps $t = 0, 1, 2, 3...$ Refer *Figure.1*. At each time t, the agent observes the environment state s_t , and chooses an action a_t , which causes the environment to transition to a new state s_{t+1}, and to reward the agent with r_{t+1} . In a Markovian system, the next state and reward depend only on the current state and present action taken, in a stochastic manner. To clarify notation used below, in a system with discrete number of states, S is the set of states. Likewise, A is the set of all possible actions and $A(s)$ is the set of actions available in states. The objective of the agent is to learn to maximize the expected value of reward received over time. It does this by learning a (possibly stochastic) mapping from states to actions called a policy. More precisely, the objective is to choose each action at so as to maximize the expected return R, given by,

$$R := E\left(\sum_{i=0}^{\infty} \gamma^i r_{t+i+1}\right) \qquad (1)$$

where γ is the discount-rate parameter in range [0,1] , which allows the agent to trade-off between the immediate reward and future possible rewards.

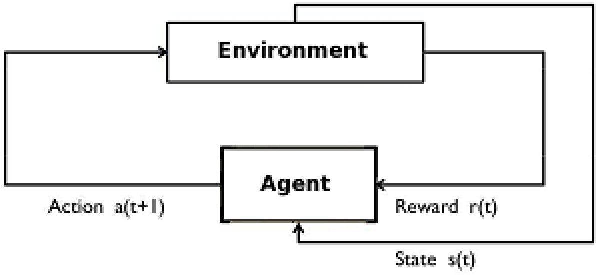


Fig.1 Concept of Reinforcement learning depicting iteraction between agent and environment

Two common solution strategies for learning an optimal policy are to approximate the optimal value function, V*, or the optimal action-value function, Q*. The optimal value function maps each state to the maximum expected return that can be obtained starting in that state and thereafter always taking the best actions. With the optimal value function and knowledge of the probable consequences of each action from each state, the agent can choose an optimal policy. For control problems where the consequences of each action are not necessarily known, a related strategy is to approximate Q*, which maps each state and action to the maximum expected return starting from the given state, assuming that the specified action is taken, and that optimal actions are chosen thereafter. Both V* and Q* can be defined using *Bellman -equations* as

$$Q^*(s,a) = \sum_{s' \epsilon S} P_{ss'}^{a} \left[R_{ss'}^{a} + \gamma \max_{a' \epsilon A(s')} Q^*(s',a')\right] \qquad (2)$$

where s' is the state at next time step, $P^{a}_{ss'}$ is its probability of transission and $R^{a}_{ss'}$ is the associated reward.

4. OUR APPROACH

4.1. Problem formulation

In this paper, we want to study the application of machine-learning in operating systems and build learning modules so as to make the timeslice parameter flexible and adaptive. Our aim is to maintain the generality of our program so that it can be employed and learned in any environment. We also want to analyze how long it takes for a module to learn from its own experiences so that it can be usefully harnessed to save time. Our main approach is to employ reinforcement learning techniques for addressing this issue of continuous improvement. We want to formulate our learning through the reward-function which can self-evaluate its performance and improve overtime.

Our prime motivation is to reduce the redundant preemptions which current schedulers do not take into account. To explain using a simple example, suppose a process has a very little burst time left and it is swapped due to the completion of its timeslice ticks, then the overhead of cache-invalidation, pipeline clearing, context switching etc. reduces the efficiency. Hence having a flexible timeslice window will prevent the above scenario. This would also improve the total time taken after the submission of process to its completion, in-return creating more processor ticks for future.

4.2. Module Design

Figure.2 gives an over all view of our entire system. It describes how our reinforcement learning agent makes use of the patterns learned initially and later on after having enough experiences it develops a policy of itself to use the prior history and reward-function.

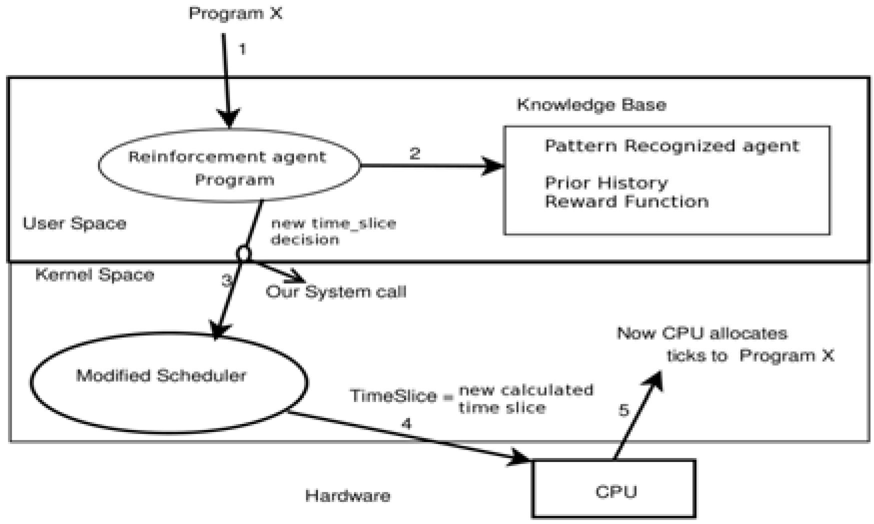


Fig.2 Bird's eye view of our design and implementation pipeline

Formally, these below steps capture the important end-to-end flow mechanism.

1. Program X passes its requirements in user-space for acquiring resources from computer hardware. These requirements are received by our agent.

2. Reinforcement learning agent uses its knowledge base to make decision. It uses patterns recognized in the initial stages to have a kick start with reasonable values and not random values. Later on knowledge base develops its history and reward function after sufficient number of experiences.

3. The information is passed from the user-space to kernel-space via a system call which will have to be coded by us. This kernel call will redirect the resource request to our modified scheduler.

4. The number of ticks to be allocated is found in the fields of new_timeslice and forwarded to CPU. And finally, CPU allocates these received orders in form of new ticks.

As the intermediate system call and modified scheduler are the only changes required in the existing systems, we provide complete abstraction to the CPU and user-space.

4.3. Modelling an RL agent

We present here a model to simulate and understand the Reinforcement learning concepts and understand the updates of Bellman equation in greater depth [6]. We have created this software with an aim to visualize the results of changing certain parameters of RL functions and as a precursor for modelling scheduler.

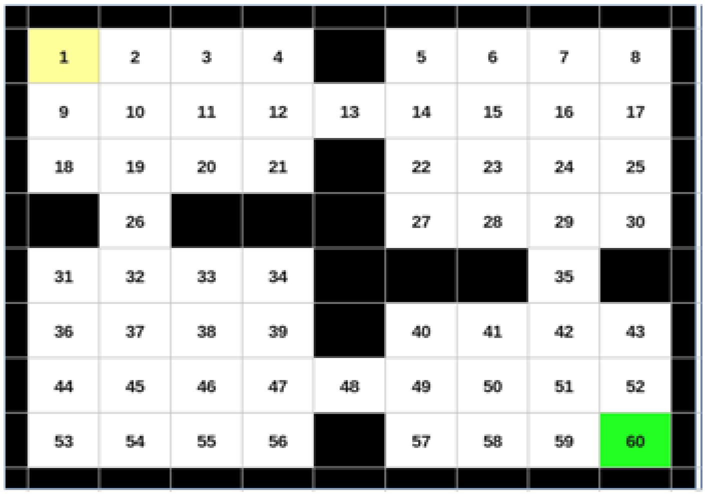


Fig.3 Maze showing the environment in which RL-agent interacts.

- *Work-Space*: Checkerboard setup with a grid like maze.

- *Aim*: To design an agent which finds its own way from the start state to goal state. The agent is designed to explore the possible paths from start state and arrive at goal state. Each state has four possible actions N, S, E & W. Collision with wall has no effect.

- *Description*: *Figure.3* depicts the maze which consists of rooms and walls. The whole arena is broken into states. Walls are depicted by dark-black solid blocks denoting that the agent cannot occupy these positions. The other blocks are number 1,2,3.....60 as the possible states in which agent can be. Agent is situated at S_1 at time t=0 and at every future action it tries to find its way to the goal state S_{60}.

- *Reward-function*: Transition to goal state gives a reward of +100. Living penalty is 0. Hence the agent can take as long time as it wants to learn the optimal policy. This parameter will be changed in case of real time schedulers. *Reward Updating policy* has Temporal difference updates with learning rate (alpha) =0.25

Initially the agent is not aware of its environment and explores it to find out. Later it learns a policy to make that wise decision about its path finding. Code (made publicly available) is written in C language for faster excution time and the output is an HTML file to help better visualize the reward updates and policy learned. Results and policies learned will be described in later sections.

4.4. Simulation

As the scheduler resides deep in the kernel, measuring the efficacy of scheduling policies in Linux is difficult. Tracing can actually change the behavior of scheduler and hide defects or inefficiencies. For example, an invalid memory reference in the scheduler will almost certainly crash the system [8]. Debugging information is limited and not easily obtained or understood by new developer. This combination of long crash-reboot cycles and limited debugging information can result in a time-consuming development process. Hence we resort to a good simulator of the Linux scheduler which we can manipulate for verifying our experiments instead of changing kernel directly.

LinSched: Linux Scheduler simulation

LinSched is a Linux scheduler simulator that resides in user space [11]. It isolates the scheduler subsystem and builds enough of the kernel environment around it that it can be executed within user space. Its behaviour can be understood by collecting relevant data through a set of available APIs. Its wrappers and simulation engine source is actually a Linux distribution. As LinSched uses the Linux scheduler within its simulation, it is much simpler to make changes, and then integrate them back into the kernel.

We would like to mention few of the essential simulator side APIs below, which we experimented over. One can utilize them to emulate the system calls and program the tasks. They are used to test any policy which are under development and see the results beforehand implementing at kernel directly. *linsched_create_RTrr(...)* -creates a normal task and sets the scheduling policy to normal. *void linsched_run_sim(...)* -begins a simulation. It accepts as its only argument the number of ticks to run. At each tick, the potential for a scheduling decision is made and returns when it is complete. Few statistics commands like *void linsched_print_task_ stats()* and *void linsched_print_group_stats()* give more detailed analysis about a task we use. We can find the total execution time for the task (with *task_exec_time(task)*), time spent not running (*task->sched_info.run_delay*), and the number of times the scheduler invoked the task (*task->sched_info.pcount*).

We conducted several experiments over the simulator on normal batch of jobs by supplying it work load in terms of process creation. First 2 normal tasks are created with no difference and ambiguity (using *linsched_create_normal_task(...)*). We next created a job which runs on normal scheduler and has a higher priority by assigning *nice value* as -5. Similarly we experimented with jobs which had lower priority of +5, followed by populating another normal and neutral priority. On the other hand, we also verified our experiments over batch tasks which are created with low and high priorities. They are all computation intensive tasks which run in blocks or batches. (using *linsched_create_batch_task(...)*). And then finally one real-time FIFO task with priority varying in range of 50-90, and one round-robin real-time task with similar priority range. Each task as created is assigned with task_id which is realistic as in real linux machines. Initially all tasks are created one after other and then after *scheduler_tick()* function times out, it is called for taking decision on other processes in waiting/ready queue. The relevant results will be discussed in subsequent sections.

5. IMPLEMENTATION AND EXPERIMENTS

5.1. Knowledge Base Creation

5.1.1. Creating Dataset

To characterize the program execution behaviour, we needed to find the static and dynamic characteristics. We used readelf and size commands to get the attributes. We built the data set of approximately 80 execution instances of five programs: matrix multiplication, quick sort, merge sort, heap sort and a recursive Fibonacci number generator. For instance, a script ran matrix multiplication program of size 700 x 700 multiple times with different nice values and selected the special time slice (STS), which gave minimum Turn Around Time (TaT). After collecting the data for the above programs with different input sizes, all of them were mapped to the best priority value. Data of the above 84 instances of the five programs were then classified into 11 categories based on the attribute time slice classes with each class having an interval of 50 ticks.

We mapped the variance of timeslice against total Turnaround Time (TaT) taken by various processes like Insertion sort, Merge sort, Quick sort, Heap sorts and Matrix multiplication with input ranging from 1e4, 1e5, 1e6 after experimenting against all possible timeslices.

5.1.2. Processing Dataset

After extracting the features from executable filles, by readelf and size commands, we refine the number of attributes to only those few essential features which actually help in taking decision. A few significant deciding features which were later used for building decision tree are: *RoData* (read only data), *Hash* (size of hash table), *Bss* (size of uninitialized memory), *DynSym* (size of dynamic linking table), *StrTab* (size of string table). The less varying / non-deciding features are discarded. The best ranked special time slices to each instance to gauge were classified to the corresponding output of decision tree. The processed result was further fed as input to to the classifier algorithm (decision trees in our case) to build iterative if-else condition.

5.1.3. Classification of Data

To handle new incoming we have built a classifier with attributes obtained from previous steps. Decision tree rules are generated as the output from classification algorithm. We used WEKA (Knowledge analysis tool) to model these classifiers. Most important identified features are

RoData , *Bss* and *Hash* . Finally groups are classified into 20 classes in ranges of timeslice. Few instances for Decision Tree Rules are mentioned below.

- *if* {(RoData<=72) AND (bss <= 36000032) AND (bss <= 4800032) AND (bss <=3200032) } *then* class=**13**
- *if* {(RoData<=72) AND (bss <= 36000032) AND (bss <= 4800032) AND (bss > 3200032) } *then* class=**2**
- *if* {(RoData<=72) AND (bss <= 36000032) AND (bss > 4800032) AND (bss <= 7300032) } *then* class=**5**
- *if* {(RoData<=72) AND (bss <= 36000032) AND (bss > 4800032) AND (bss > 7300032) AND (bss <= 2000032) } *then* class=**3**
- *if* {(RoData<=72) AND (bss > 36000032) AND (bss <= 4800032) } *then* class=**7**
- *if* {(RoData<=72) AND (bss <= 36000032) AND (bss > 4800032) AND (bss > 7300032) AND (bss > 2000032) } *then* class=**0**
- *if* {(RoData<=72) AND (bss > 36000032) AND (bss <= 4800032) } *then* class=**4**

To give a better visualization of out features, we present in *Table.1* various statistics obtained for Heap sort with input size 3e5 and priority (nice value) set to 4. These statistics help us decide the lowest Turnaround Time and lowest number of swaps taken for best priority class.

Table.1 Statistics obtained for Heap sort with input size 3e5 showing the classifier features.

Feature Name	Value	Feature Name	Value
User time (seconds)	0.37	Voluntary context switches	1
Minor (reclaiming frame) page faults	743	Involuntary context switches	41
Percent of CPU this job got	98%	File system outputs	8
Elapsed (wall clock) time (h:mm:ss)	0:00.38	Socket messages sent	0
Maximum resident set size (kbytes)	1632	Socket messages received	0
Signals delivered	0	Page size (bytes)	4096

5.2. Self-Improving Module

The self-learning module which is based on Reinforcement learning technique is proved to improve over time with its experience until converged to saturation. The input to this module is the group decision from the knowledge base in the previous step as the output of the if-else clause. Further, reinforcement learning module may give a new class if it decides from its policy learned over time of several running experiences. In the background this self improving module would explore for new classes which it could assign to a new incoming process. We modelled the scheduler actions as a markov decision process where decisions for assigning a new time slice solely based over current state and it need not have to take into account of the previous decisions. The policy mapping for states and their aggregate reward associated is done using the Bellman equations

$$V^*(s) = \max_{a\in A} \sum_{s' \epsilon S} P^a_{ss'} \, [R^a_{ss'} + \gamma V^*(s')] \tag{3}$$

Figure.4 shows how using an array implementation of Doubly Linked List, we generated the above module. Temporal difference (TD method) was used for updating the reward-function with experiences and time. The sense of reward for the scheduler agent was set to be a function of inverse of waiting time of the process. The choice of such a reward function was to avoid the bias introduced by the inverse of total turnaround time (TaT), which is the least where compared to waiting time. This is because TaT is also inclusive of total number of swaps which in turn is dependent over the size of input and size of text, whereas waiting time does not depend over the size of input. We set the exploration vs. exploitation constant to be 0.2 which is still flexible under temperature coefficient mentioned above.

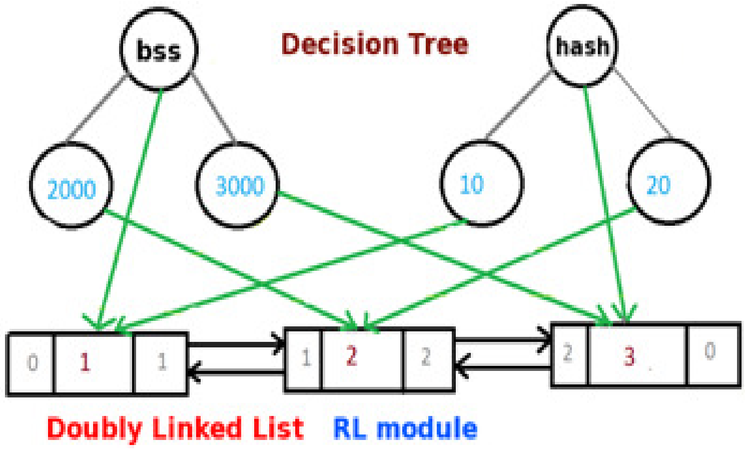


Fig.4 Integration of self learning module with decision tree knowledge base.

Input to this module is the class decision from knowledge base obtained in the previous step which is the output of the decision tree. It outputs a new class which RL module decides from its policy generated over time of running. Reward sense is given by the inverse of waiting time of the process. We have used exploration vs. exploitation ε-greedy constant as 0.2.

In our experimental Setup, we used WEKA (Knowledge analysis tool) for Decision trees and attribute selection. For compilation of all programs we used gcc (GNU_GCC) 4.5.1 (RedHat 4.5.1-4). To extract the attributes from executable/binary we used readelf & size command tools. For graph plots and mathematical calculations we used Octave.

6. Results And Analysis

Below we present a few test cases which characterise the general behaviour of scheduler interaction with knowledge base and self improving module. We also analyse the effectiveness of integrating Static knowledge base and self-learning module by calculating time saved and number of CPU cycles conserved. Programs were verfified after executing multiple times with different nice values on Linux System. Their corresponding figures show how the turn-around-time changed as the CPU allotted timeslice of the process changed.

Experiments show that there does not exist any direct evident relation between time slice and CPU utilization performance metrics. Refer *Figure.5* and *Figure.6* plot of TaT vs. timeslice class allotted. Hence it is not a simple linear function which is monotonic in nature. One will have to learn a proper classifier which can learn the pattern and predict optimal timeslice. Below we show the analysis for 900x900 matrix multiplcation and merge sort (input size 3e6). *Table.2* shows their new suggested class from knowledge base. For Heapsort (input size 6e5) and Quicksort (input size 1e6) we have only plotted their TaT vs. Timeslice graphs in *Figure.6*, which is similar in wavy nature as *Figure.5*. We have omitted explicit calculations to prevent redundancy in paper, as their nature is very similar to previous matrix multiplication.

Effectiveness analysis for Matrix Multiplication with input size of 900x900 random matrix elements.

- Turnaround Time (normal) - 27872216 ms
- Turnaround Time (with KB)- 24905490 ms
- Time saved = 2966726 ms
- Time saved per second - 109879 ms
- No. of clock cycles saved - 2.4MHz x 109879
- No. of Lower operations saved - 109879 / (pipeline clear + context switch etc.)

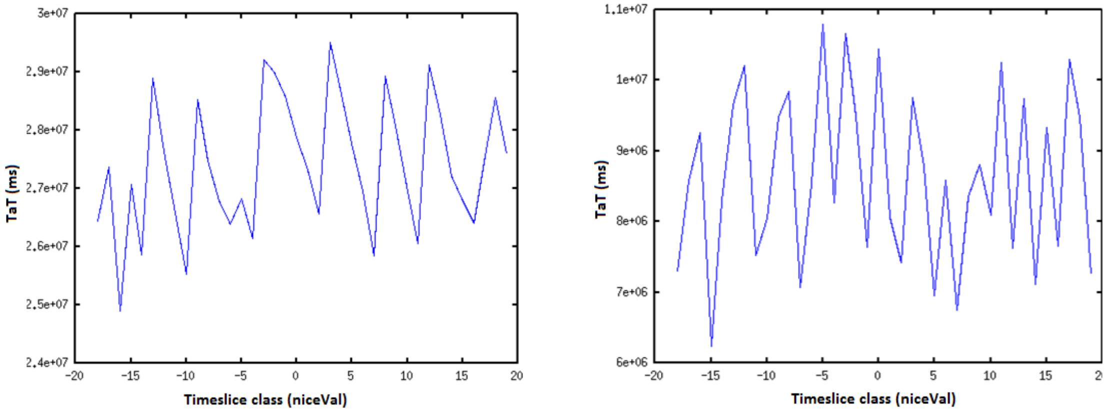


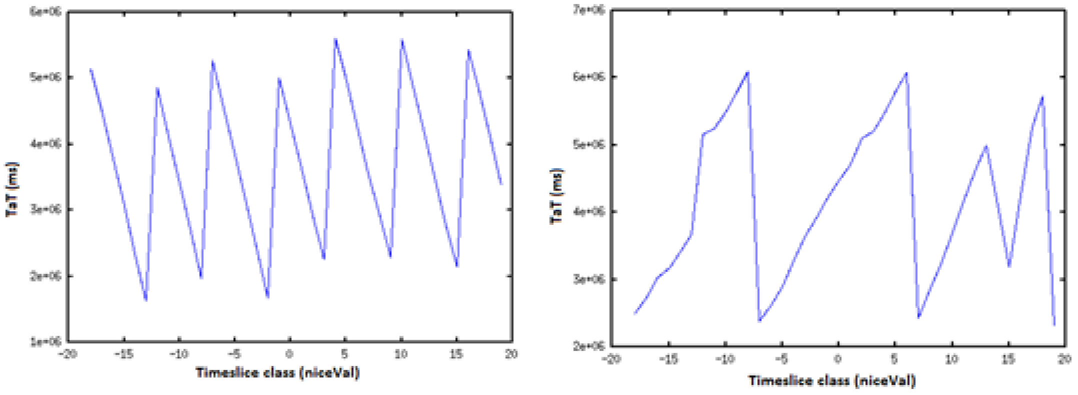
Fig.5 Timeslice class (unnormalized nice values) vs. Turn around time for (a)Matrix Multiplication and (b)Merge Sort .

Table.2 Optimal timeslice-class decisions made by knowledge base module for Matrix multiplication of input 900x900 and Merge sort over input size 3e6 elements.

Matrix Multiplcation		Merge Sort	
Turn around time (microsec)	Timeslice class suggested	Turn around time (microsec)	Timeslice class suggested
24905490	16	6236901	15
25529649	10	7067141	7
25872445	14	7305964	18
26151444	4	7524301	11
26396064	6	7639436	1
26442902	18	8055423	10
26577882	11	8273131	4
26800116	7	8302461	14
26827546	5	8537245	6
27080158	15	8569818	17
27376257	17	9255897	16
27484162	8	9483901	9
27643193	12	9499732	2
28535686	9	9660585	13
28581739	1	9844913	8
28900769	13	10217774	12

Effectiveness analysis for Merge Sort with input size of 3e6 random array elements.

- TaT (normal) - 10439931 ms and TaT(with KB)- 6236901 ms
- Time saved = 4203030 ms
- Time saved per second - 382093 ms
- No. of clock cycles saved - 2.4MHz x 382093
- No. of Lower operations saved - 382093 / (pipeline clear + context switch etc.)



(a) (b)

Fig.6 Timeslice class (unnormalized nice values) vs. Turn around time for (a)Heap Sort, (b)Quick Sort .

7. CONCLUSION

From the results we can observe that the turnaround time can be optimized by reducing redundant context switches and also reducing the additional lower level register swaps, pipeline clearances etc. This in turn saves the CPU cycles which are valuable resource for runtime execution of subsequent jobs. A self-learning module proposed here has the potential of constantly improving with more experiences and is provided over a knowledge base to prevent the problem of cold-start. We have showed the non-intuitive irregularity between decreasing turnaround time and increasing time slice by wave-pattern of TaT vs. class of time. The machine learning module identifies specific time slice for given task so that its resource usage and TaT are optimized. We are also currently investigating ways to address the problem of infinite horizon in reinforcement learning, as the scheduler may run for infinite amount of time (or very large time unit) and scores rewards just for the sake of its existence. The agent may actually be performing sub-optimally, but its prolonged existence keeps collecting rewards and show positive results. This issue can be addressed by refreshing the scheduler after certain time unit, but clarity is required over how to calculate its optimal refresh period.

REFERENCES

[1] Warren Smith, Valerie Taylor, Ian Foster, Predicting Application Run-Times Using Historical Information", Job Scheduling Strategies for Parallel Processing, IPPS/SPDP'98 Workshop, March, 1998.

[2] Jonathan M Eastep, "Smart data structures: An online ML approach to multicore Data structure", IEEE Real-Time and Embedded Technology and Applications Symposium 2011.

[3] M. John Calandrino , documentation for the main source code for LinSched and author the linux_linsched files LINK: http://www.cs.unc.edu/~jmc/linsched/

[4] D. Vengerov, A reinforcement learning approach to dynamic resource scheduling allocation, Engineering Applications of Artificial Intelligence, vol. 20, no 3, p. 383-390, Elsevier, 2007.

[5] Richard Gibbons, A Historical Application Profiler for Use by Parallel Schedulers, Lecture Notes on Computer Science, Volume : 1297, pp: 58-75, 1997.

[6] Richard S. Sutton and Andrew G. Barto. , Reinforcement Learning: An Introduction. A Bradford Book. The MIT Press Cambridge, Massachusetts London, England.

[7] Atul Negi, Kishore Kumar. P, UOHYD, Applying machine learning techniques to improve Linux process scheduling 2010.

[8] Internals of Linux kernel and documentation for interface modules LINK:
http://www.faqs.org/docs/kernel_2_4/lki-2.html

[9] D. Vengerov, A reinforcement learning framework for utility-based scheduling in resource-constrained systems, Future Generation Compute Systems, vol. 25, p. 728-736 Elsevier, 2009.

[10] Surkanya Suranauwarat, Hide Taniguchi, The Design, Implementation and Initial Evaluation of An Advanced Knowledge-based Process Scheduler, ACM SIGOPS Operating Systems Review, volume: 35, pp: 61-81, October, 2001.

[11] Documentation for IBM project for real scheduler simulator in User space LINK: http://www.ibm.com/developerworks/library/l-linux-schedulersimulator/

[12] Tong Li, Jessica C. Young, John M. Calandrino, Dan P. Baumberger, and Scott Hahn , LinSched: The Linux Scheduler Simulator Research Paper by Systems Technology Lab Intel Corporation, 2008

[13] McGovern, A., Moss, E., and Barto, A. G. (2002). Building a basic block instruction scheduler with reinforcement learning and rollouts. Machine Learning, 49(2/3):141– 160.

[14] Martin Stolle and Doina Precup , Learning Options in Reinforcement Learning Springer-Verlag Berlin Heidelberg 2002

[15] Danie P. Bovet, Marc, Understanding the Linux Kernel, 2nd ed, O' Reilly and Associates, Dec., 2002.

[16] Modern Operation System Scheduler Simulator, development work for simulating LINK: http://www.ontko.com/moss/

[17] Andrew Marks, A Dynamically Adaptive CPU Scheduler, Department of Computer Science, Santa Clara University, pp :5- 9, June, 2003.

AUTHORS

Prakhar Ojha, student of the Department of Computer Science Engineering at National Institute of Technology Karnataka Surathkal, India. His areas of interest are Artificial Intelligence, Reinforcement Learning and Application of knowledge bases for smart decision making.

Siddhartha R Thota, student of the Department of Information Technology at National Institute of Technology Karnataka Surathkal, India. His areas of interest are Machine Learning, Natural language processing and hidden markov model based Speech processing.

Vani M is a Associate Professor in the Computer Science Engineering Department of NITK. She has over 18 years of teaching experience. Her research interest includes Algorithmic graph theory, Operating systems and Algorithms for wireless sensor networks.

Mohit P Tahiliani is a Assistant Professor in the Computer Science Engineering Department of NITK. His research interest includes Named Data Networks, TCP Congestion Control, Bufferbloat, Active Queue Management (AQM) mechanisms and Routing Protocol Design and Engineering.

A NEW APPROACH BASED ON THE DETECTION OF OPINION BY SENTIWORDNET FOR AUTOMATIC TEXT SUMMARIES BY EXTRACTION

Reda Mohamed HAMOU[1], Mohamed Amine BOUDIA[2] and Abdelmalek AMINE[3]

Department of Computer, Dr. Moulay Tahar University Saida, Algeria
Laboratory Knowledge Management and Complex Data (GeCoDe Lab) Saida, Algeria
{hamoureda1, mamiamounti2, abd_amine1 3}@yahoo.fr

ABSTRACT

In this paper, we propose a new approach based on the detection of opinion by the SentiWordNet for the production of text summarization by using the scoring extraction technique adapted to detecting of opinion. The texts are decomposed into sentences then represented by a vector of scores of opinion of this sentences. The summary will be done by elimination of sentences whose opinion is different from the original text. This difference is expressed by a threshold opinion. The following hypothesis: "textual units that do not share the same opinion of the text are ideas used for the development or comparison and their absences have no vocation to reach the semantics of the abstract" Has been verified by the statistical measure of Chi_2 which we used it to calculate a dependence between the unit textual and the text. Finally we found an opinion threshold interval which generate the optimal assessments.

KEYWORDS

Automatic Summary Extraction, Text Mining, Evaluation, Automatic Language Processing, F-Measure, correlation, ROUGE-SU (2), SentiWordNet, Opinion Mining.

1. INTRODUCTION AND PROBLEMATIC

Currently, one of the major problems for computer scientists is access to the content of information, access itself or in other words the software and hardware infrastructure are no longer an obstacle, and the major problem is the exponential increase in the amount of textual information electronically. This requires the use of more specific tools i.e. access to the content of texts by rapid and effective means has become a necessary task.

A summary of a text is an effective way to represent the contents of the texts and allow quick access to their semantic content. The purpose of a summarization is to produce an abridged text covering most of the content from the source text.

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 163–181, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51514

Summary of text appears interesting for fast access to the content of textual information. A summary is a reissued the original text in smaller form that is done under the constraint of keeping the semantics of a document that is minimized entropy semantics. The purpose of this operation is to help the reader identify interesting information for him without having to read the entire document. The uses of automatic summaries aim to reduce the time to find the relevant documents or reduce treatment long texts by identifying the key information. The volume of electronic textual information is increasing, making access to information difficult. Producing a summary may facilitate access to information, but it is also a complex task because it requires language skills.

To do an automatic summarization, the current literature presents three approaches:

- Automatic Summarization by extraction
- Automatic Summarization by understanding
- Automatic Summarization by automatic classification

Another line of research that has gained momentum in recent years, in case the Opinion Mining or the fact of detecting opinion of a sentence, paragraph or text. Our job is to use detection methods to produce a summary opinion. We propose the hypothesis:

"Textual units that do not share the same opinion of the text are ideas used for the development or comparison and their absences have no vocation to reach the semantics of the abstract"

In this work we will generate a summary automatically by extraction approach, we will use the scoring technique where the score will calculate according to opinion by using a SentiWordNet

We will build a summary of the sentences that have an opinion similar to that of full text according to a threshold of opinion; our work will give an answer for the following question:

- Have our hypothesis been testable? If so, is it valid?
- What is the impact of opinion threshold on the quality of the summary?
- The opinion mining can he bring a plus for automatic summarization?

2. LITERATURE REVIEW

Automatically produce a summary is an idea that has emerged in the early 1950s, this is a branch of natural language processing (NLP). The first attempts made their apparition in 1950, the community has tried to implement simple approaches as extracting relevant sentences according to a scoring in order to arrive at a summary understandable and easily readable by a human. (Luhn, 1958) [1] (Edmundson, 1969)[2] proposed to identify lexical units carriers of semantics by manual analysis, the result is called extracts (in English), i.e. an extract is a summary built by phrases (considered as pertinent) original text. Their idea is to assign a weight to each sentence that represents its pertinence then extract either by a reduced rate, N sentences whose weight are greater, or a threshold scoring, tell that all the sentences with a score and greater than or equal to the threshold will be kept.

Other work focuses on automating the analysis for the detection of pertinent lexical units: several methods and variations have been proposed (Radev et al, 2001)[3]; (Radev et al, 2004)[4]; (Boudin and al , 2007)[5]; (Carbonell and al, 1998)[6].(Boudin and al, 2008)[7] we see that we can select two sentences pertinent but if it looks like, they have worked on eliminating redundancy by similarity measures sentences, they even prove their method detects the topic text.

The proposed approaches are intuitive, since the 1960s (Edmundson, 1969)[2] pro-poses to make an identification of keywords (based on a theme), followed by a job that involves a consideration of the position of sentences in the document. The MEAD2 system is the most popular nowadays, developed instead by Ramdev, he implemented the approach of this type (Radev et al., 2001)[3]. He identified the most salient words in each text, which he calls "centroid", and A extract is made by sen-tence who contain the greatest number of "centroid". Another Neo-cortex system was developed at the University of Avignon, which is based on the combination of screening measures, phrases to select the benefits covered by each of them (action selection), this system obtained very good results by evaluating the algorithm MMR (Maximal Marginal Relevance) of (Carbonell & Goldstein, 1998)[6].

Some years later other approaches were presented, these approaches are based on knowledge representation (Mani, 2004)[8], the thematic segmentation (Farzindar et al, 2004)[9] Or recognition of user profiles (author of full text) (Crispino & Couto, 2004)[10].

Noting some attempts to introduce deep linguistic systems in automatic summarization, most recent work was on the syntax tree by providing a method for comparing between syntactic trees, make the elimination of syntax trees or merge (Barzilay & McKeown, 2005)[11];. This attempt was followed by another job that offers alternative methods of syntactic compression based on theoretical and empirical linguistic properties (Yousfi-Monod, 2007)[12].

Most jobs are not interested in the opinion contained in the text, at the end of 2000 the community began to have more specialized query summaries, including dealing with the detection and analysis of the opinion. (Eyrich, et al, 2001)[13] propose a system diagram that has not been implemented: This system integrates a QR module and an analysis module opinion to make a summary of the responses to the opinion without changing the content.

Work on automatic summaries have neglected early analysis of the opinion. Analysis of opinion is divided into three main levels of subtasks: the first sub-task is to distinguish between a subjective texts and an objective text (Yu & Hatzivassiloglou, 2003) [14]. The second sub-task is to classify texts subjective positive or negative (Turney, 2002)[15]. The third level of refinement trying to determine the extent to which positive or negative texts (Wilson et al., 2004)[16]. The impulse given by campaigns such as TREC Blog task opinion since 2006 is undeniable (Pang & Lee, 2008)[17]. (Zhang et al, 2007)[18];(Dey & Haque, 2009)[19].

Opinion Mining is an area that has attracted many researchers which resulted in several works. There are two types of approaches for detecting opinion: Approaches based on corpus (Corpus-based Approach) (Hatzivassiloglou and McKeown, 1997)[20] (Wiebe, 2000) [21] (Kanayama and Nasukawa, 2006) [22] (Esuli and Sebastiani 2006) [23]; and (Qiu et al, 2009) [24], others based on a dictionary (Dictionary-based Approach) dictionary (Hu and Liu, 2004) [25], (Kim and Hovy, 2004) [26], (Kamps et al, 2004) [27], (Esuli and Sebastiani , 2005)[28], (Andreevskaya and Bergler, 2006) [29], and (Bouchlaghem et al, 2010) [30].

This is the second approach that will be used in this article, (Wu and Liu, 2004) use the adjectives for the detection of opinions. They manually build a list of adjectives they use to predict the orientation of the sentence and use WordNet to supply the list synonyms and antonyms of adjectives whose polarity is known. In (Liu and al, 2007)[31], the authors count the number of occurrences of each entity in the "Pros" expressing a positive opinion and "Cons" that negative opinions. In (Zhang and Liu, 2011) [32], the authors showed that the noun and noun phrases can also enclose opinions, They count the number of positive and negative sentences for each feature of the product using the lexicon of opinion prepared by (Ding and al, 2008)[33].Strength (intensity) of opinion is also required; Indeed, subjectivity is expressed in different ways; "Good battery" is different from "great battery" and "excellent battery." (Pang and Lee, 2008)[17] focus on detecting the strength of opinion using the techniques of boosting, rule learning and support vector regression. (Pang and Lee, 2008) [17]and (Turney, 2002)[15] classify documents as "thumbs up" or "thumbs down" according to the opinion they convey. However, (Pang and Lee, 2005)[34] exploit machine learning techniques to give a score from 1 to 5 on passages opinions. While (Esuli and Sebastiani 2006)[28] construct the SentiWordNet that a dictionary of general opinion; currently in its third version SentiWordNet 3.0; SentiWordNet can be defined as a lexical resource designed specifically for use by application of detecting opinions and feelings. SentiWordNet is the result of an annotation of all synsets of WordNet so that it assigns to each word (synset) an opinion score. SentiWordNet contains 1000 synsets which makes it very small compared with WordNet, besides the 1000 synsets SentiWordNet automatically ignores all other inputs. Another weakness is that several synsets are not carrying opinion[35].

we can not pretend that our opinion detection for the production of summary text goes beyond the assessment of degrees of positivity or negativity, we must shed light on recent efforts to introduce more linguistic and discourse approaches (taking into account the modality of the speaker) in this accompli by (Asher and al., 2008)[36].

As for the evaluation of abstracts is a crucial problem, emphasize the contribution (Goulet, 2007)[37] that goes beyond the coverage of n-grams and offers a terminology adapted to French. In recent years, large-scale assessments, independent designers systems have emerged and several evaluation measures have been proposed. As regards the assessment of automatic summary, two evaluation campaigns have already been conducted by the U.S. DARPA (Defense Advanced Research Projects Agency). The first, entitled SUMMAC, ran from 1996 to 1998 under the TIPSTER (Lin and al., 2003) [38] program, and the second, entitled DUC (Document Understanding Conferences) (noting that France still lagging behind several countries in all science especially: Computer science), (Das and al., 2007) [39] followed from 2000 to 2007. Since 2008 it is the Text Analysis Conference Such measures may be applied to the distribution of units in the summaries of P systems and those of reference Q. The method was evaluated by Lin et al. (2006) on the corpus DUC'02 for tasks mono and multi-document summary. A good correlation was found between measures of divergence and the two rankings obtained with ROUGE and coverage. (Louis & Nenkova ,2009) [39] went further and, proposed to compare the distribution of words in the complete documents with the words in automatic summaries to infer an evaluation measure based on the content.

3. OUR PROPOSED APPROACH

Our approach is based on the identification of opinions textual units (phrases, clauses, sentences, paragraphs), the identification of opinion original text, finally extracting textual units that share the same opinion that the original text

We start from the hypothesis mentioned in previous section.

Recall that the opinion is an expression of the feelings of a person towards an enti-ty or an aspect of the entity (Liu, 2010). An entity may be a product, a person, event, organization or topic.

SentiWordNet is used to filter the word bearer of opinion first, then we will use the score returned by SentiWordNet like so:

```
      If (score_opinion (term i) <0) then the opinion of (term i)
is negative, else opinion is positive
```

Our approach will follow the following steps:

3.1 Pretreatment

Simple cleaning: Empty words will not be deleted, because the method for automatic summarization by extraction is based on extracting the most informative sentences without change and because the final result is a text (abstract) : if any words will be deleted without information on their morphosyntactic and semantic impact in sentences, you can get a text summary inconsistent.

And for this cleaning will be limited to delete emoticons and replace spaces with _ and remove the special characters that cannot fit in French or English literature (#, \, [,]...........)

Choice of term: for automatic summarization by extraction we will need two representations:

- Bag of words representation.
- Bag sentence representation.

The two representations are introduced in the context of vector model:

The first representation is to transform the text into a vector vi (w_1 , w_2 , ... , $w_{|T|}$) where T is the number of all the words appearing at least once in the text. The weight w_k indicates the occurrence of the word t_k in the document.

The representation is to transform the text into a vector V^1_i (w'_1 , w'_2 , ... , $w'_{|R|}$) where R is the number of all the phrases that appear at least once in the text. The w'_k weight indicates the occurrence of t_k sentence in the document.

And finally a matrix of occurrence sentences * will generate a word from the two previous representations, the size of the matrix is :

- The number of words in the text * the number of words in the text,
- The weight pik represents the number of occurrences of the word k in sentence i;

3.2 Detecting opinion by SentiWordNet:

The "sentence-term" matrix is reduced to a "sentence - carrier term of opinion " matrix filtering the term vector vi by SentiWordNet, no-existing terms in the dictionary opinion will be eliminated.

At the end of this step, a matrix M of size nxp where n is equal to the number of phrases and p is equal to the number of term carrier opinion. M_{ij} indicates the occurrence of the word (opinion holder) j in sentence i

$$O_{ij} = M_{ij} * score\ (j) \qquad (1)$$

The score (j) is the score obtained by the SentiWordNet for the term j

3.3 Construction of Summary

Weighting: Once the matrix "Phrase- carrier term of opinion" is ready, we calculate the score of phrases as well as the score of text in order to proceed to the final step.

The opinion score for textual units (sentences, paragraph or text) is equal to average the score of the holders of opinion obtained by the SentiWordNet terms.

So the score of opinion of each sentence will be calculated as follows:

$$Score_{phrase}[i] = \frac{\sum_{i=0}^{n} O_{ij}}{\sum_{i=0}^{n} M_{ij}} \qquad (2)$$

such that n = number of carrier term of opinion in the textual unit

Finally, we identified the opinion of text that is the average of opinion score of phrase that the compound:

$$Score_{texte} = \frac{\sum_{i=0}^{R} Score_{phrase}[i]}{R} \qquad (3)$$

As size R of vector V' (number of sentences)

Summary Final: "The suggested procedure claims on the principle that high-frequency words in a document are important words" [Luhn 1958]. In our case we will adapt this quote as follows: "The suggested procedure claims on the principle that sentences that share the same opinion that the document (text) are important phrases" ie phrases that do not share the same opinion with text without phrases that have been used by the author to develop an idea or comparison is equivalent to saying that we can eliminate them without causing a large entropy of sense.

The final step is to select the phrases that have the same opinions as the text, for this we proposed this method:

By neighborhood threshold of opinion of phrases: we kept the phrase his degree of similarity of opinion between this phrase and the text is greater than or equal threshold neighborhood or opinion threshold.

For each sentence k do
If (score_texte - threshold_opinion < threshold _phrase [k]) and (threshold _phrase [k] <score_texte + threshold _opinion)
Then selected the sentence k

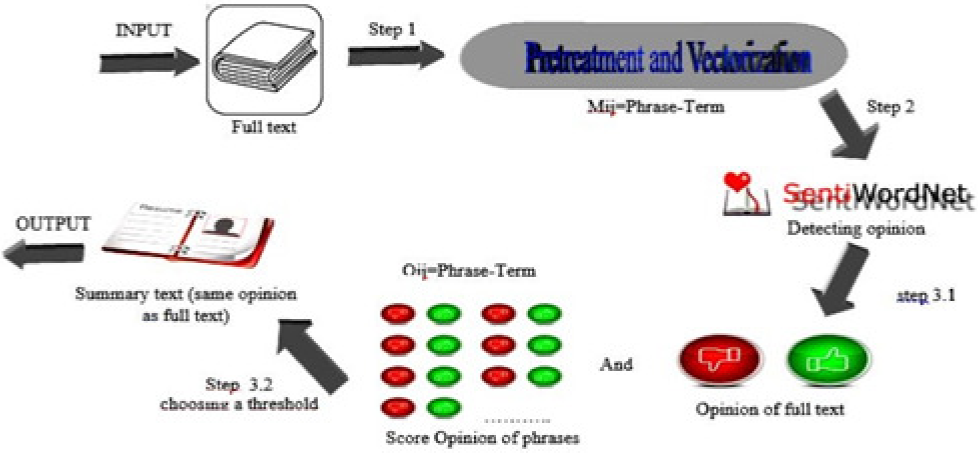


Fig. 1. Full process of the proposed approach

4. EXPERIMENTATION

To test our hypothesis already mentioned we use the Chi2 measure is a well-known statistical measure, it evaluates the lack of independence between a textual unit and a text. It uses the same concepts of co-occurrence word / text mutual information, but the difference lies on standardization, which makes them comparable terms. Measurement Chi_2 still loses relevance for infrequent terms.[41]

$$X^2\,(ut_k, text_i) = \frac{|T|.\,[P\,(ut_k, text_i).\,P\,(\overline{ut_k}, \overline{text_i}) - \,P\,(ut_k, text_i).\,P\,(\overline{ut_k}, text_i)]^2}{P(ut_k).\,P(\overline{ut_k}).\,P(text_i).\,P(\overline{text_i})} \quad (4)$$

The use of measurement Chi_2 determines the independence of sentences eliminated by detecting opinion with the original text. Chi_2 promotes the absence of terms and the most common and takes into consideration information from the text terms. A high value of the Chi-2 (k, i) reflects a dependency between the sentence k and the text i .

The second step of our experiment will begin to study is a robustness of summary, we use two evaluation method ROUGE-SU (2) (Recall-Oriented Understudy for Gisting Evaluation –Skip Unit) and F-measure.

4.1 Used corpus

Was used as the text corpus "Hurricane" in the English text (to use the SentiWordNet): the text contains a title and 20 sentences.

Summaries reference: we took three reference summary product successively by Summarizer CORTEX, Essentiel Summarizerand the third by a human expert.

Cortex

Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas. The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to feet to Puerto Rico's south coast.

Essential Summarizer

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.

Human expert

Hurricane Gilbert is moving toward the Dominican Republic, where the residents if the south coast, especially the Barahona Province, have been alerted to prepare for heavy rain, and high winds and seas. Tropical storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2am Sanday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

4.2 Validation

To estimate the robustness of our summary, we calculate the correlation metric ROUGE-SU (2) (Lin 2004) which compares a candidate summary (automatically generated by the system to be evaluated) and a reference summary (created by human experts or other automatic summarization systems known). And another F-Measure metric that we have proposed in earlier work.

The evaluation measure Recall - Oriented Understudy for Gisting Evaluation.

The evaluation of abstracts can be done semi-automatically through measures of similarities computed between a candidate summary and one or more reference summaries. We evaluate the results of this work by the measure called Recall - Oriented Understudy for Gisting Evaluation

(ROUGE) proposed by (Lin, 2004) involving the differences between distributions of words. Heavily used in the DUC campaigns, these measures are increasingly considered standard by the community because of their strong correlation with manual ratings. Two variants of ROUGE will be developed; it is the measures used in the DUC campaigns.

ROUGE (N)

Measurement of return calculated on the co-occurrences of N-grams between a candidate summary R_{can} and a set of reference summaries R_{ref} Co-occurrences (N-gram) is the maximum number of co-occurrences N-grams and R_{ref} in R_{can} number and (N-grams) to the number of N-grams appearing in the abstract

$$ROUGE\ (N) = \frac{\sum_{s\in R_{ref}} \sum_{s\in R_{can}} \mathrm{Co-occurences}\ (R_{ref}, R_{can}, N)}{\mathrm{Nbr-NGramme}\ (N)_{R_{ref}}} (5)$$

ROUGE-SU (M)

Adaptation of ROUGE-2 using bigrams hole (skip units (SU)) maximum size M and counting unigrams.

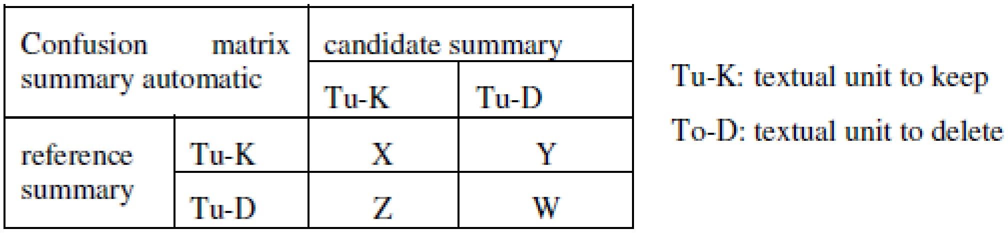
F-Measure pour l’évaluation des résumés automatique par extraction

We proposed in our previous work an adaptation of the F-measure for the validation of automatic summarization by extraction, since this technique is based on sentences to keep and delete else following some a philosophy (scoring, detection Thematic....), this can be considered as a two-class classification.

Our method is a hybrid between the two valuation methods: intrinsic and extrinsic.

We shall compare the applicant and the full text (summary) summary to identify textual units that have been kept and which have been deleted, then did the same operation between the reference summary and the full text (summary), and finally a comparison between the two summaries (candidate and reference) is performed to obtain the following confusion matrix.

Table 1. Confusion matrix summary automatic



Confusion matrix summary automatic		candidate summary	
		Tu-K	Tu-D
reference summary	Tu-K	X	Y
	Tu-D	Z	W

Tu-K: textual unit to keep

To-D: textual unit to delete

A Recall of "Tu-K" class is defined by the number of textual units kept in the candidate and reference summary (shared), divided by the number of units of text kept by the reference summary; in parallel calculates Recall of "Tu-D" class in the same way that is to say, the number of text units deleted in the candidate and reference summary (shared), divided by the number of units of text deleted reference summary

$$Rappel_{\mathrm{Tu-K}} = \frac{X}{X+Y} (6) \quad Rappel_{\mathrm{To-D}} = \frac{W}{W+Z} (7)$$

The precision of "Tu-K" class is defined by the number of textual units kept in the candidate and reference summary (shared), divided by the number of units of textual kept by the candidate summary; in parallel calculates the precision class "Tu-D" in the same way that is to say, the number of text units removed from the candidate and reference summary (shared), divided by the number of textual units deleted by the candidate summary.

$$Précision_{\mathrm{Tu-K}} = \frac{X}{X+Z}(8) \qquad Précision_{\mathrm{To-D}} = \frac{W}{W+Y}(9)$$

Since automatic summarization by extraction is a two-class classification thus:

$$R\mathrm{appel} = \frac{\mathrm{Rappel_{Ut\ toG}}+\mathrm{Rappel_{Ut\ toS}}}{2}(10) \qquad \mathrm{Précision} = \frac{\mathrm{Précision_{Ut\ toG}}+\mathrm{Précision_{Ut\ toS}}}{2}(11)$$

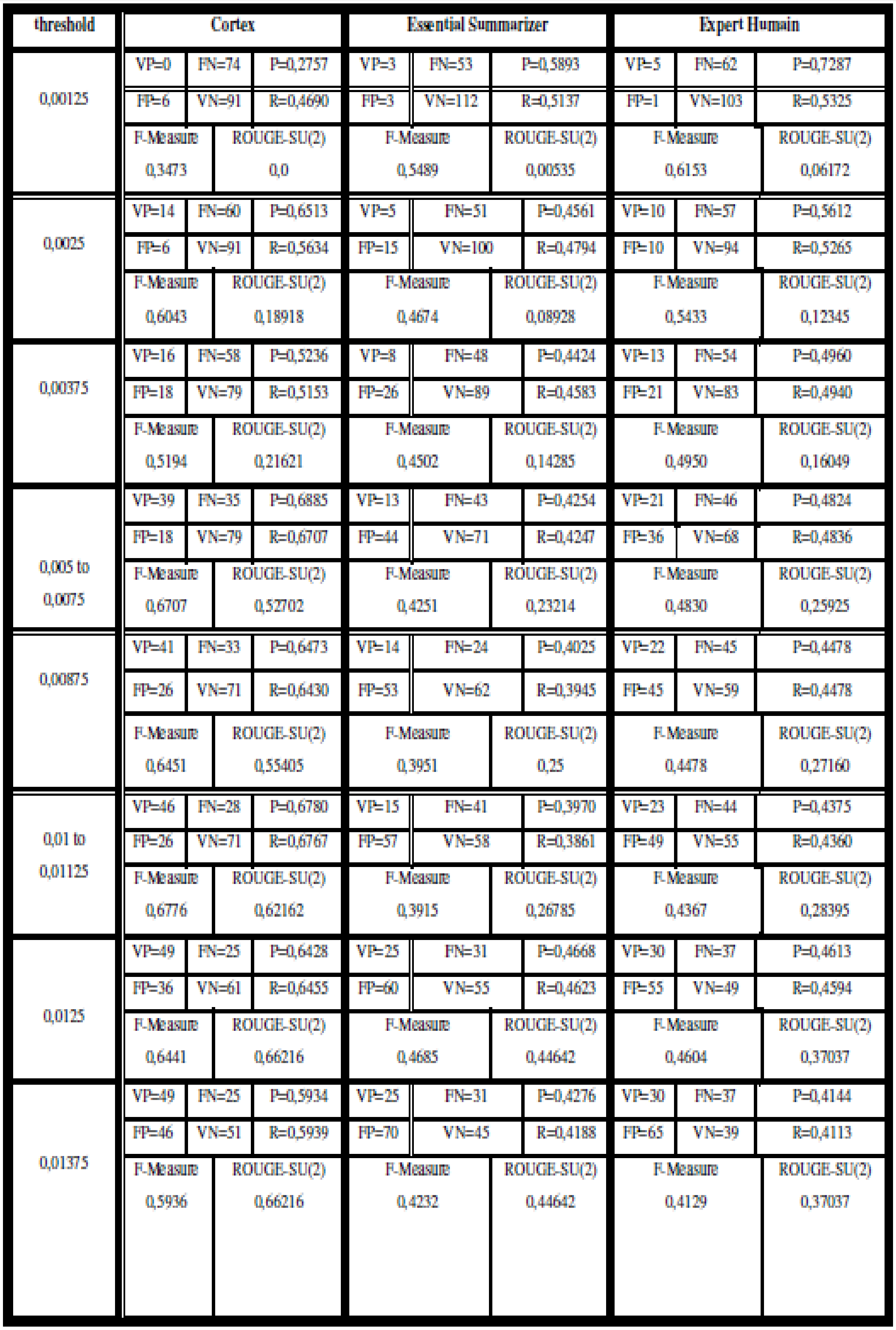
Finally combining precision and recall is calculated weighting to the F-Measure

$$F - Mesure = \frac{2 * (Précision * Rappel)}{(Précision + Rappel)}(12)$$

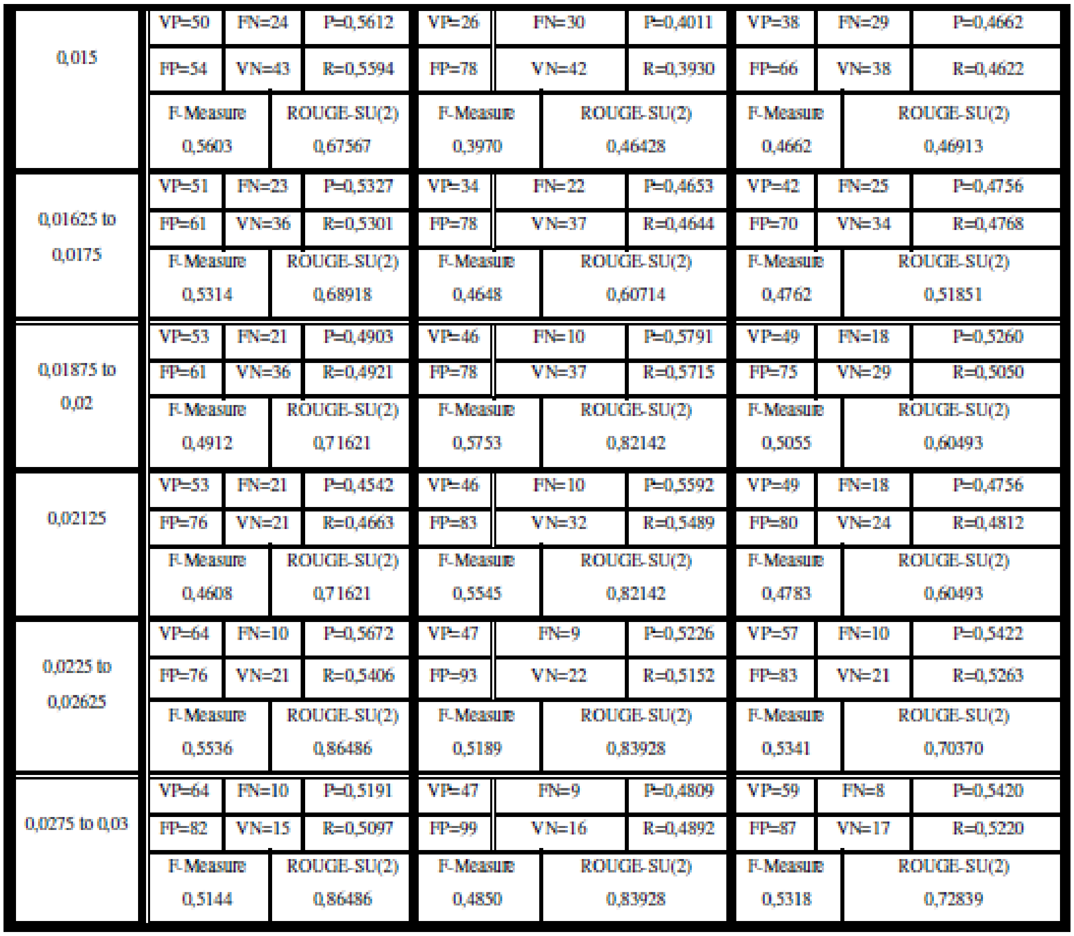
4.3 The algorithm description.

```
Begin
Pretreatment paper
Mij = integer array [1 .. number of sentence] [1 .. number of terms]
Mij = vectorization (Word Bag, bag of sentences)
Oij = integer array [1 .. number of sentence] [1 .. number of terms]
For each sentence i do begin
For each j terms to begin
                    $O_{ij} = M_{ij} * score\ (j)(13)$
End for
End for
Score _phrase = array real [1 .. number of sentence]
for each sentence k do begin
                    $Score\_phrase\ [i] = \frac{\sum_{i=0}^{n} O_{ij}}{\sum_{i=0}^{n} M_{ij}}(14)$
End for
                    $Score\_texte = \frac{\sum_{i=0}^{R} Score_{phrase}[i]}{R}(15)$
Threshold = défini_par _l'utilisateur
For each sentence do i start
If (score_texte-seuil_voisingane <score_phrase [i] <+ score_texte seuil_voisingane)
then remember sentence [i] if not eliminated sentence [i]
End for
END
```

4.4 Result



threshold	Cortex			Essential Summarizer			Expert Humain		
0,00125	VP=0	FN=74	P=0,2757	VP=3	FN=53	P=0,5893	VP=5	FN=62	P=0,7287
	FP=6	VN=91	R=0,4690	FP=3	VN=112	R=0,5137	FP=1	VN=103	R=0,5325
	F-Measure 0,3473		ROUGE-SU(2) 0,0	F-Measure 0,5489		ROUGE-SU(2) 0,00535	F-Measure 0,6153		ROUGE-SU(2) 0,06172
0,0025	VP=14	FN=60	P=0,6513	VP=5	FN=51	P=0,4561	VP=10	FN=57	P=0,5612
	FP=6	VN=91	R=0,5634	FP=15	VN=100	R=0,4794	FP=10	VN=94	R=0,5265
	F-Measure 0,6043		ROUGE-SU(2) 0,18918	F-Measure 0,4674		ROUGE-SU(2) 0,08928	F-Measure 0,5433		ROUGE-SU(2) 0,12345
0,00375	VP=16	FN=58	P=0,5236	VP=8	FN=48	P=0,4424	VP=13	FN=54	P=0,4960
	FP=18	VN=79	R=0,5153	FP=26	VN=89	R=0,4583	FP=21	VN=83	R=0,4940
	F-Measure 0,5194		ROUGE-SU(2) 0,21621	F-Measure 0,4502		ROUGE-SU(2) 0,14285	F-Measure 0,4950		ROUGE-SU(2) 0,16049
0,005 to 0,0075	VP=39	FN=35	P=0,6885	VP=13	FN=43	P=0,4254	VP=21	FN=46	P=0,4824
	FP=18	VN=79	R=0,6707	FP=44	VN=71	R=0,4247	FP=36	VN=68	R=0,4836
	F-Measure 0,6707		ROUGE-SU(2) 0,52702	F-Measure 0,4251		ROUGE-SU(2) 0,23214	F-Measure 0,4830		ROUGE-SU(2) 0,25925
0,00875	VP=41	FN=33	P=0,6473	VP=14	FN=24	P=0,4025	VP=22	FN=45	P=0,4478
	FP=26	VN=71	R=0,6430	FP=53	VN=62	R=0,3945	FP=45	VN=59	R=0,4478
	F-Measure 0,6451		ROUGE-SU(2) 0,55405	F-Measure 0,3951		ROUGE-SU(2) 0,25	F-Measure 0,4478		ROUGE-SU(2) 0,27160
0,01 to 0,01125	VP=46	FN=28	P=0,6780	VP=15	FN=41	P=0,3970	VP=23	FN=44	P=0,4375
	FP=26	VN=71	R=0,6767	FP=57	VN=58	R=0,3861	FP=49	VN=55	R=0,4360
	F-Measure 0,6776		ROUGE-SU(2) 0,62162	F-Measure 0,3915		ROUGE-SU(2) 0,26785	F-Measure 0,4367		ROUGE-SU(2) 0,28395
0,0125	VP=49	FN=25	P=0,6428	VP=25	FN=31	P=0,4668	VP=30	FN=37	P=0,4613
	FP=36	VN=61	R=0,6455	FP=60	VN=55	R=0,4623	FP=55	VN=49	R=0,4594
	F-Measure 0,6441		ROUGE-SU(2) 0,66216	F-Measure 0,4685		ROUGE-SU(2) 0,44642	F-Measure 0,4604		ROUGE-SU(2) 0,37037
0,01375	VP=49	FN=25	P=0,5934	VP=25	FN=31	P=0,4276	VP=30	FN=37	P=0,4144
	FP=46	VN=51	R=0,5939	FP=70	VN=45	R=0,4188	FP=65	VN=39	R=0,4113
	F-Measure 0,5936		ROUGE-SU(2) 0,66216	F-Measure 0,4232		ROUGE-SU(2) 0,44642	F-Measure 0,4129		ROUGE-SU(2) 0,37037

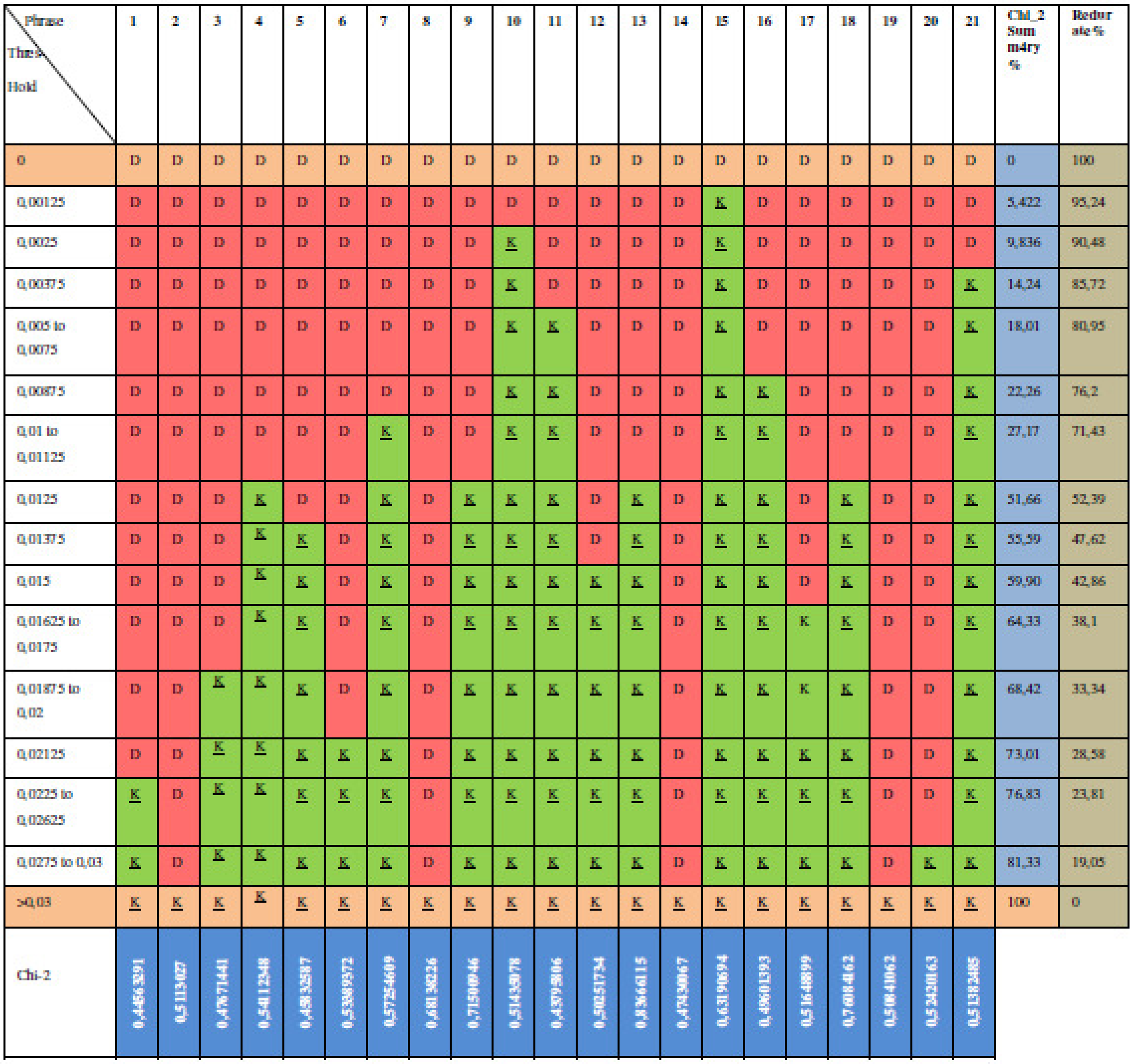


0,015	VP=50	FN=24	P=0,5612	VP=26	FN=30	P=0,4011	VP=38	FN=29	P=0,4662
	FP=54	VN=43	R=0,5594	FP=78	VN=42	R=0,3930	FP=66	VN=38	R=0,4622
	F-Measure 0,5603	ROUGE-SU(2) 0,67567		F-Measure 0,3970	ROUGE-SU(2) 0,46428		F-Measure 0,4662	ROUGE-SU(2) 0,46913	
0,01625 to 0,0175	VP=51	FN=23	P=0,5327	VP=34	FN=22	P=0,4653	VP=42	FN=25	P=0,4756
	FP=61	VN=36	R=0,5301	FP=78	VN=37	R=0,4644	FP=70	VN=34	R=0,4768
	F-Measure 0,5314	ROUGE-SU(2) 0,68918		F-Measure 0,4648	ROUGE-SU(2) 0,60714		F-Measure 0,4762	ROUGE-SU(2) 0,51851	
0,01875 to 0,02	VP=53	FN=21	P=0,4903	VP=46	FN=10	P=0,5791	VP=49	FN=18	P=0,5260
	FP=61	VN=36	R=0,4921	FP=78	VN=37	R=0,5715	FP=75	VN=29	R=0,5050
	F-Measure 0,4912	ROUGE-SU(2) 0,71621		F-Measure 0,5753	ROUGE-SU(2) 0,82142		F-Measure 0,5055	ROUGE-SU(2) 0,60493	
0,02125	VP=53	FN=21	P=0,4542	VP=46	FN=10	P=0,5592	VP=49	FN=18	P=0,4756
	FP=76	VN=21	R=0,4663	FP=83	VN=32	R=0,5489	FP=80	VN=24	R=0,4812
	F-Measure 0,4608	ROUGE-SU(2) 0,71621		F-Measure 0,5545	ROUGE-SU(2) 0,82142		F-Measure 0,4783	ROUGE-SU(2) 0,60493	
0,0225 to 0,02625	VP=64	FN=10	P=0,5672	VP=47	FN=9	P=0,5226	VP=57	FN=10	P=0,5422
	FP=76	VN=21	R=0,5406	FP=93	VN=22	R=0,5152	FP=83	VN=21	R=0,5263
	F-Measure 0,5536	ROUGE-SU(2) 0,86486		F-Measure 0,5189	ROUGE-SU(2) 0,83928		F-Measure 0,5341	ROUGE-SU(2) 0,70370	
0,0275 to 0,03	VP=64	FN=10	P=0,5191	VP=47	FN=9	P=0,4809	VP=59	FN=8	P=0,5420
	FP=82	VN=15	R=0,5097	FP=99	VN=16	R=0,4892	FP=87	VN=17	R=0,5220
	F-Measure 0,5144	ROUGE-SU(2) 0,86486		F-Measure 0,4850	ROUGE-SU(2) 0,83928		F-Measure 0,5318	ROUGE-SU(2) 0,72839	

Table 2. Result Evaluation of summary produced (candidate) with ROUGE and F-measure using 3 reference summary Cortex, Essential Summaries, human expert (second part)

The above table includes an assessment summary with a different threshold of opinions comparing with all reference summary using the F - measure and ROUGE. The following table shows in a manner explicit the selected sentences (keep) (K) and the sentences deleted (D) at each threshold of opinion and gives the chi_2 value for each sentence with original text and the chi2 rate of each summary report by the original text.

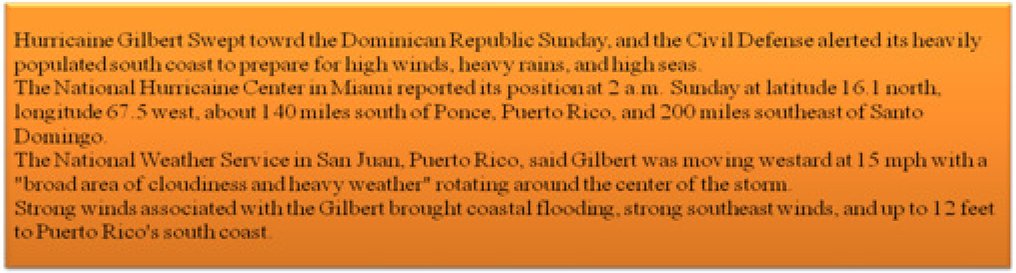
The rate of chi-2 which is equal to the sum of value chi_2 sentence divided by retained by the number of all the phrases which constitutes the original text, it indicates the correctness of choice of phrases relative to their dependence original text.



Phrase Thres Hold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Chi_2 Sum m4ry %	Redur ate %
0	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	0	100
0,00125	D	D	D	D	D	D	D	D	D	D	D	D	D	D	K	D	D	D	D	D	D	5,422	95,24
0,0025	D	D	D	D	D	D	D	D	D	K	D	D	D	D	K	D	D	D	D	D	D	9,836	90,48
0,00375	D	D	D	D	D	D	D	D	D	K	D	D	D	D	K	D	D	D	D	D	K	14,24	85,72
0,005 to 0,0075	D	D	D	D	D	D	D	D	D	K	K	D	D	D	K	D	D	D	D	D	K	18,01	80,95
0,00875	D	D	D	D	D	D	D	D	D	K	K	D	D	D	K	K	D	D	D	D	K	22,26	76,2
0,01 to 0,01125	D	D	D	D	D	D	K	D	D	K	K	D	D	D	K	K	D	D	D	D	K	27,17	71,43
0,0125	D	D	D	K	D	D	K	D	K	K	K	D	K	D	K	K	D	K	D	D	K	51,66	52,39
0,01375	D	D	D	K	K	D	K	D	K	K	K	D	K	D	K	K	D	K	D	D	K	55,59	47,62
0,015	D	D	D	K	K	D	K	D	K	K	K	K	K	D	K	K	D	K	D	D	K	59,90	42,86
0,01625 to 0,0175	D	D	D	K	K	D	K	D	K	K	K	K	K	D	K	K	K	K	D	D	K	64,33	38,1
0,01875 to 0,02	D	D	K	K	K	D	K	D	K	K	K	K	K	D	K	K	K	K	D	D	K	68,42	33,34
0,02125	D	D	K	K	K	K	K	D	K	K	K	K	K	D	K	K	K	K	D	D	K	73,01	28,58
0,0225 to 0,02625	K	D	K	K	K	K	K	D	K	K	K	K	K	D	K	K	K	K	D	D	K	76,83	23,81
0,0275 to 0,03	K	D	K	K	K	K	K	D	K	K	K	K	K	D	K	K	K	K	D	K	K	81,33	19,05
>0,03	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	100	0
Chi-2	0,44563291	0,5113027	0,47671441	0,54112348	0,45832587	0,53889372	0,57254609	0,68138226	0,71500946	0,51435078	0,43793806	0,50251734	0,83666115	0,47430067	0,63190694	0,49601393	0,51648899	0,76084162	0,50841062	0,52420163	0,51382485		

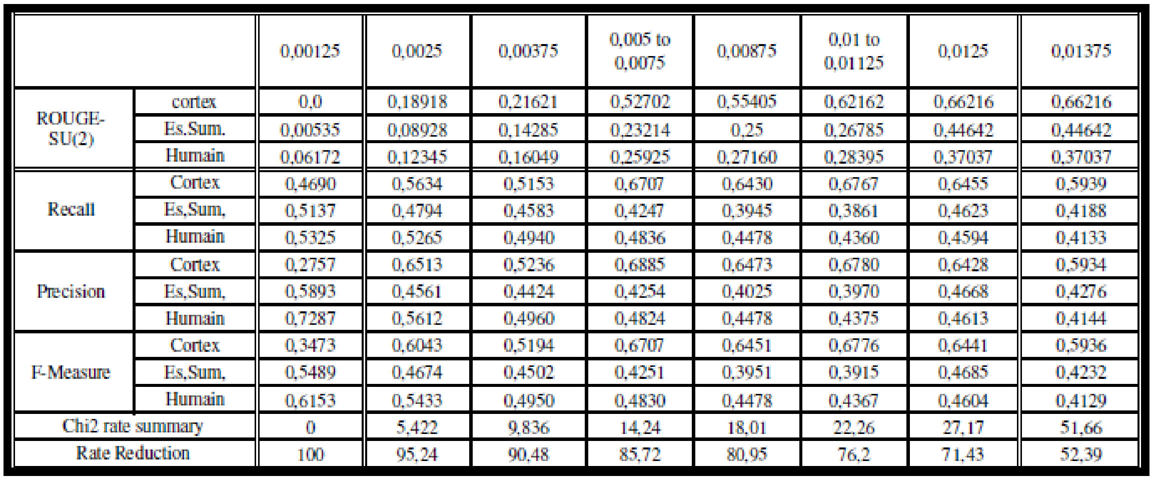
Table 3. Distribution of sentences in summary product (candidate) for each threshold of opinion , his chi_2 rate and reduction rate, and Chi_2 value for each sentence

4.5 The best summary text



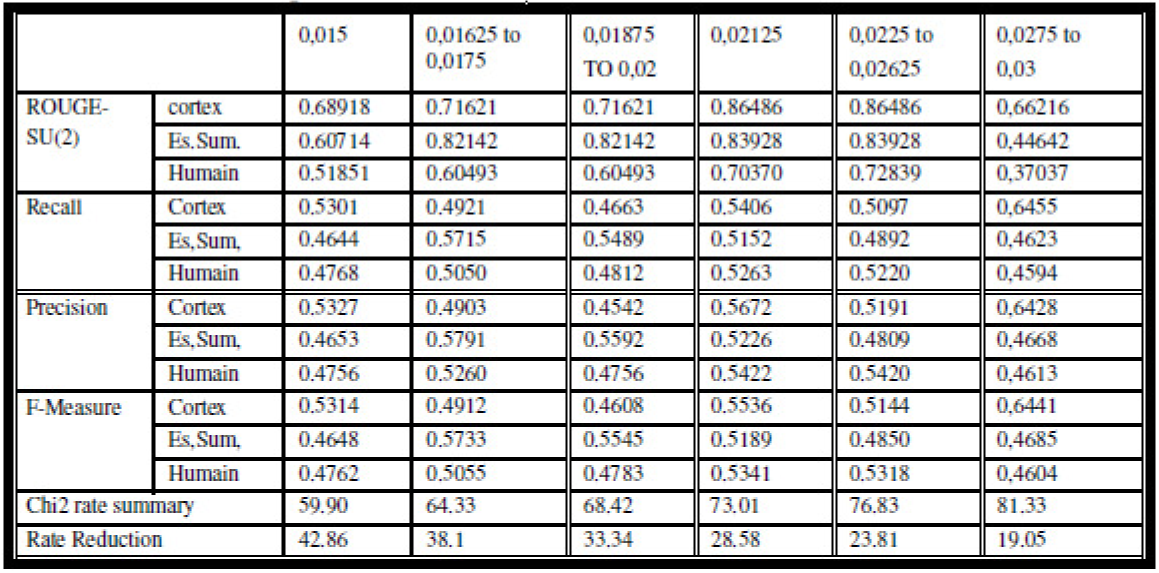
Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.

The following table summarizes all the results mentioned in the two previous tables in case the table2 and table 3:



		0,00125	0,0025	0,00375	0,005 to 0,0075	0,00875	0,01 to 0,01125	0,0125	0,01375
ROUGE-SU(2)	cortex	0,0	0,18918	0,21621	0,52702	0,55405	0,62162	0,66216	0,66216
	Es.Sum.	0,00535	0,08928	0,14285	0,23214	0,25	0,26785	0,44642	0,44642
	Humain	0,06172	0,12345	0,16049	0,25925	0,27160	0,28395	0,37037	0,37037
Recall	Cortex	0,4690	0,5634	0,5153	0,6707	0,6430	0,6767	0,6455	0,5939
	Es,Sum,	0,5137	0,4794	0,4583	0,4247	0,3945	0,3861	0,4623	0,4188
	Humain	0,5325	0,5265	0,4940	0,4836	0,4478	0,4360	0,4594	0,4133
Precision	Cortex	0,2757	0,6513	0,5236	0,6885	0,6473	0,6780	0,6428	0,5934
	Es,Sum,	0,5893	0,4561	0,4424	0,4254	0,4025	0,3970	0,4668	0,4276
	Humain	0,7287	0,5612	0,4960	0,4824	0,4478	0,4375	0,4613	0,4144
F-Measure	Cortex	0,3473	0,6043	0,5194	0,6707	0,6451	0,6776	0,6441	0,5936
	Es,Sum,	0,5489	0,4674	0,4502	0,4251	0,3951	0,3915	0,4685	0,4232
	Humain	0,6153	0,5433	0,4950	0,4830	0,4478	0,4367	0,4604	0,4129
Chi2 rate summary		0	5,422	9,836	14,24	18,01	22,26	27,17	51,66
Rate Reduction		100	95,24	90,48	85,72	80,95	76,2	71,43	52,39

this is the second part of Table above



		0,015	0,01625 to 0,0175	0,01875 TO 0,02	0,02125	0,0225 to 0,02625	0,0275 to 0,03
ROUGE-SU(2)	cortex	0.68918	0.71621	0.71621	0.86486	0.86486	0,66216
	Es.Sum.	0.60714	0.82142	0.82142	0.83928	0.83928	0,44642
	Humain	0.51851	0.60493	0.60493	0.70370	0.72839	0,37037
Recall	Cortex	0.5301	0.4921	0.4663	0.5406	0.5097	0,6455
	Es,Sum,	0.4644	0.5715	0.5489	0.5152	0.4892	0,4623
	Humain	0.4768	0.5050	0.4812	0.5263	0.5220	0,4594
Precision	Cortex	0.5327	0.4903	0.4542	0.5672	0.5191	0,6428
	Es,Sum,	0.4653	0.5791	0.5592	0.5226	0.4809	0,4668
	Humain	0.4756	0.5260	0.4756	0.5422	0.5420	0,4613
F-Measure	Cortex	0.5314	0.4912	0.4608	0.5536	0.5144	0,6441
	Es,Sum,	0.4648	0.5733	0.5545	0.5189	0.4850	0,4685
	Humain	0.4762	0.5055	0.4783	0.5341	0.5318	0,4604
Chi2 rate summary		59.90	64.33	68.42	73.01	76.83	81.33
Rate Reduction		42.86	38.1	33.34	28.58	23.81	19.05

Table 4. Recapitulation of table 1 and 2, ROUGE-SU(2) vs F-Measure, Vs chi_2 rate summary Vs rate reduction

5. INTERPRETATION

We tested our approach with an incremental threshold 0.00125 to see the impact of threshold of opinion on the quality of the summary and to recommend range threshold that returns good results.

ROUGE is a intrinsic semi-automatic evaluation metric based on the number of co-occurrence between a candidate summary and one or more reference summaries divided by the size of the latter. Its weakness is that it is based on references summary and neglects the original text.

The value given by ROUGE for a summary with a negligible reduction rate is high. This high value is explained by the to the increased number of co-occurrence between the candidate summary and references summary.

The F-Measure is one of the most robust metric and most used for the evaluation of classification; The F-measure is a combination of Recall and precision. For our adaptation we added to the force F-Measures an extrinsic evaluation in the beginning, and continue with an intrinsic evaluation: So this is hybrid evaluation. For automatic summary reduced rate of reduction, F-measure gives better than ROUGE assessments because it takes account of the absence of term. But unlike ROUGE, evaluating a candidate summary with high reduction ratio summary can be distorted, because FALSE NEGATIVE FN and TRUE NEGATIVE TN is the maximum which will give good result in summary generally poor (highest reduction rate leads to an increase of entropy information)

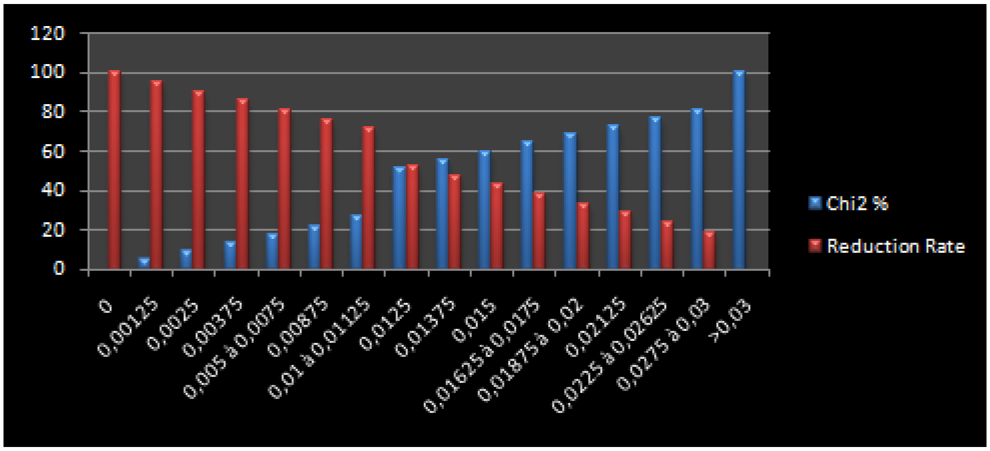


Fig. 2. Chi_2 rate candidate summary Vs reduction rate candidate summary

The precision indicated the purity of the candidate summary, while recall interprets the likeness of the candidate summary reference.

We can deduce from the above graph that the two variable: rate reduction rate Chi_2 text and have an inverse correlation, indicating that the increased number of selected sentences increases the independence original text. This indication is logic and expected, returning to table 2, we can see that in the made to retain P4, P 13 ,P18 sentences we increase the rate of chi-2 + by 25%, thanks to their strong dependence on text that is equal to 0.57 for the P4 and larger than 0.71 for P13 and P18

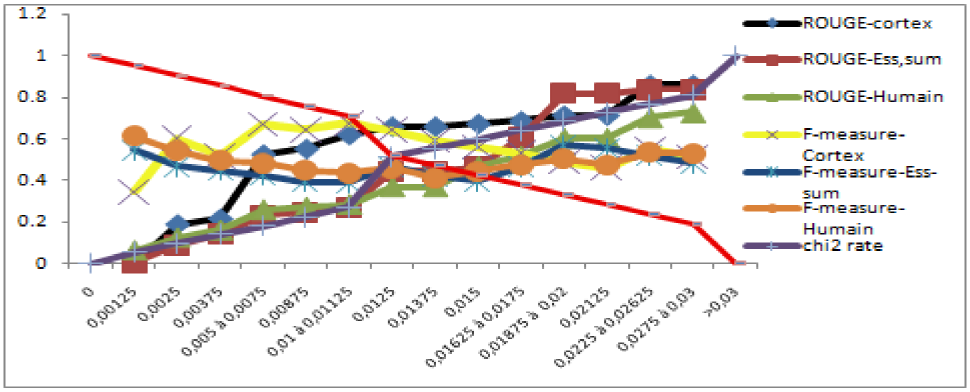


Fig. 3. ROUGE VS F-Measure (with 3 reference summary) vs chi2 rate vs reduction rate

We see the ROUGE index (black curve, red and green) is higher when the rate of reduction (red curve) is low; unlike rate chi-2 indicating a great loss to the sentence dependente with a orginale text, in this case assessment ROUGE is false, and this is the result of the high value of the co-occurrence between the candidate summary and the reference summary, the probability found that of co-occurrence between a long and a short text reference is more important than a short text summary with the same reference.

Seen on the graph that the F-measure overstates the automatic summarization which has a high rate of reduction, by against it does not overstate the summary is a low rate of reduction; is explained by the consideration of missing words (False Negative: textual untités deleted in the candidate resumé but retained by the abstract and negative reference True: textual untités deleted in the candidate and the reference summary resumé). This is a strong point of the F-Measure adapted to enable automatic extraction summaries, although it must be noted that its weakness against the very small summary and True Negative achieving the maximum value (all deleted sentences in the summary Reference will also be summarized in this summary has a high reduction ratio)

Finally, we can see all indexes used for the evaluation values are reached their optimal threshold set between 0.0125 and 0.0175, in this interval all evaluation value is good for the candidate summary.

We can see from Figure 3 and Table 4 that the selection of words that are less dependent originally the text does not improve Result, for example: between 0.0125 and 0.01375 threshold, the only difference is the selection of the fifth sentence in 0.01375 (it was not in the selected 0.0125 threshold), in Table 4 we seen Chi_2 the value of this sentence is low which is also readable on the graph not stagnation the ROUGE and a slight drop in F-measure for three reference summary. This confirms our hypothesis.

6. CONCLUSION AND PERSPECTIVE

In this article, we presented a new approach for the production of an automatic summary extraction based on the detection of conscience SentiWordNet.

First line, we proposed a hypothesis that will support this approach **"textual units that do not share the same opinion of the text are ideas used for the development or comparison and their absences have no vocation to reach the semantics of the abstract "**

The second line, we explain our approach to detecting and opinion proposed a flexible technique to choose the sentence that is near to the original text opinion poll threshold.

Given the results obtained, we have validated our hypothesis; and therefore this work can help solve one of the major problems of automatic summarization: the reduction of information entropy and conservation semantics.

Looking ahead, we will try to improve automatic summarization by extraction based on the detection of opinion by the application of technical and other conventional method such as detection thematic.

7. ANNEX

Title : Hurricaine Gilbert

Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement.

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night. The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast. There were no reports on casualties. San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night. On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 90 mph winds and sheets of rain. Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

REFERENCES

[1] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

[2] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264-285.

[3] Radev, D. R., Blair-Goldensohn, S., & Zhang, Z. (2001). Experiments in single and multi-document summarization using MEAD. Ann Arbor, 1001, 48109.

[4] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 457-479.

[5] Boudin, F., & Moreno, J. M. T. (2007). NEO-CORTEX: a performant user-oriented multi-document summarization system. In Computational Linguistics and Intelligent Text Processing (pp. 551-562). Springer Berlin Heidelberg.

[6] Carbonell, J., & Goldstein, J. (1998, August). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 335-336). ACM.

[7] Boudin, F., & El-Bèze, M. (2008). A scalable MMR approach to sentence scoring for multi-document update summarization.

[8] Mani, I., Pustejovsky, J., & Sundheim, B. (2004). Introduction to the special issue on temporal information processing. ACM Transactions on Asian Language Information Processing (TALIP), 3(1), 1-10.

[9] Farzindar, A., Lapalme, G., & Desclés, J. P. (2004). Résumé de textes juridiques par identification de leur structure thématique. Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte: solutions et perspectives, 45(1), 26.

[10] CRISPINO, G., & COUTO, J. (2004). Construction automatique de résumés: Une approche dynamique: Résumé automatique de textes. TAL. Traitement automatique des langues, 45(1), 95-120.

[11] Barzilay, R., & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3), 297-328.

[12] Yousfi-Monod, M. (2007). Compression automatique ou semi-automatique de textes par élagage des constituants effaçables: une approche interactive et indépendante des corpus (Doctoral dissertation, Université Montpellier II-Sciences et Techniques du Languedoc).

[13] Eyrich, V. A., Martı-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., ... & Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics, 17(12), 1242-1243.

[14] Yu, H., & Hatzivassiloglou, V. (2003, July). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

[15] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

[16] Wilson, T., Wiebe, J., & Hwa, R. (2004, July). Just how mad are you? Finding strong and weak opinion clauses. In aaai (Vol. 4, pp. 761-769).

[17] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

[18] Zhang, Y. C., Medo, M., Ren, J., Zhou, T., Li, T., & Yang, F. (2007). Recommendation model based on opinion diffusion. EPL (Europhysics Letters), 80(6), 68003.

[19] Dey, L., & Haque, S. K. (2009, July). Studying the effects of noisy text on text mining applications. In Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data (pp. 107-114). ACM.[

[20] Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics (pp. 174-181). Association for Computational Linguistics.

[21] Wiebe, J. (2000, July). Learning subjective adjectives from corpora. In AAAI/IAAI (pp. 735-740).

[22] Kanayama, H., & Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 355-363). Association for Computational Linguistics.

[23] Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC (Vol. 6, pp. 417-422).

[24] Qiu, H., Xue, L., Ji, G., Zhou, G., Huang, X., Qu, Y., & Gao, P. (2009). Enzyme-modified nanoporous gold-based electrochemical biosensors. Biosensors and Bioelectronics, 24(10), 3014-3018.

[25] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[26] Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[27] Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004, May). Using WordNet to Measure Semantic Orientations of Adjectives. In LREC (Vol. 4, pp. 1115-1118).

[28] Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC (Vol. 6, pp. 417-422).

[29] Ohana, B., & Tierney, B. (2009, October). Sentiment classification of reviews using SentiWordNet. In 9th. IT & T Conference (p. 13).

[30] Elkhlifi, A., Bouchlaghem, R., & Faiz, R. (2011, March). Opinion Extraction and Classification Based on Semantic Similarities. In FLAIRS Conference.

[31] Liu, J., Cao, Y., Lin, C. Y., Huang, Y., & Zhou, M. (2007, June). Low-Quality Product Review Detection in Opinion Summarization. In EMNLP-CoNLL (pp. 334-342).

[32] Zhang, L., & Liu, B. (2011, June). Identifying noun product features that imply opinions. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 575-580). Association for Computational Linguistics.

[33] Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 231-240). ACM.

[34] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 115-124). Association for Computational Linguistics.

[35] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

[36] Asher, N., Benamara, F., & Mathieu, Y. Y. (2008). Distilling Opinion in Discourse: A Preliminary Study. In COLING (Posters) (pp. 7-10).

[37] Goulet, M. J. (2007). Terminologie et paramètres expérimentaux pour l'évaluation des résumés automatiques. Traitement Automatique des Langues, 48(1).

[38] Lin, C. Y., & Hovy, E. (2003, May). Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 71-78). Association for Computational Linguistics.

[39] Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU, 4, 192-195.

[40] Louis, A., & Nenkova, A. (2009, August). Automatically evaluating content selection in summarization without human models. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 306-314). Association for Computational Linguistics.

[41] Moh'd A Mesleh, A. (2007). Chi square feature extraction based SVMs Arabic language text categorization system. Journal of Computer Science, 3(6), 430-435

INTENTIONAL BLANK

A MATHEMATICAL MODEL OF ACCESS CONTROL IN BIG DATA USING CONFIDENCE INTERVAL AND DIGITAL SIGNATURE

Amine RAHMANI, Abdelmalek AMINE and Mohamed Reda HAMOU

GeCoDe Laboratory, Department of Informatics,
Dr. TAHAR Moulay university of Saida – Algeria-
Aminerahmani2091@gmail.com, amine_abd1@yahoo.fr,
hamoureda@yahoo.fr

ABSTRACT

Nowadays, the concept of big data grows incessantly; recent researches proved that 90% of the whole data existed on the web had been created in last two years. However, this growing bumped by many critical challenges resides generally in security level; the users care about how could providers protect their privacy on their data. Access control, cryptography, and de-identification are the main search areas grouped under a specific domain known as Privacy Preserving Data Publishing. In this paper, we bring in suggestion a new model for access control over big data using digital signature and confidence interval; we first introduce our work by presenting some general concepts used to build our approach then presenting the idea of this report and finally we evaluate our system by conducting several experiments and showing and discussing the results that we got.

KEYWORDS

Access control, standard deviation, privacy preserving, big data, numeric signature, confidence interval

1. INTRODUCTION

Privacy, timeless, scalability of data is the most important problems that big data recognize starting from the first step of data acquisition; in fact, one of the most disturbed principle that are used in big data is the fact of losing control on data. This concept led to a lot of criticism from clients, losing control on your own data means losing everything related to the control even the access control.

Before the coming of the concept big data, controlling access on such data was done locally using the known models such as mandatory models (MAC), discriminatory models (DAC) or role based models (RBAC) but those last cannot be used because of some impediments; in case of DAC models the users defines the right access by himself while in the use of big data the user lose the entire control on his data; in case of MAC models the right access are defined by a major

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 183–198, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51515

entity like military direction and this does not satisfy the users wishes in big data; moreover, in case of RBAC models, the right access are defines in form of roles where the can have the right from a major entity and can also give the rights on his own data to others which is bumped into reality of losing control. For that many works and propositions are passed by many researches such as in [20] and [7] using cryptography concepts and also in some of the works basing on users' identities.

In this report, we suggest a new model using some complex mathematical concepts such as standard deviation, confidence interval and primitive root to protect access control using users' identities and groups; for that we first introduce some backgrounds and definitions of the mathematical concepts that we practice, and so we introduce the main hypothesis under our good example, talking about its theoretical efficiency and carrying a set of experimentations on a set of information.

2. RELATED WORKS

The access control presents a sensitive domain in informatics security where it consists of defining such policy that allows or not for such user to get the access to such object; with the coming of concepts of big data and data sharing, this domain became a real challenge in research area. Many works are done within this highly active topic where the most of these works use a promising technique called Attribute Based Encryption such as in [32] [26] [29] and [17]; in [7] the author presented his approach of controlling hierarchical access using multiple key assignment in cryptography where he proposed four schemes, in other world four extensions of his work: bounded, unbounded, synchronous and asynchronous in order to give the general idea under temporal access control; in [2] the authors show their new approach of controlling access on resource-deprived environment in sensor data by integrating the Ladon Security Protocol that offers a secure access using end-to-end authentication, authorisation and key establishment mechanisms in PrivaKERB user privacy framework of KERBEROS environment; in [27] the authors introduced a purpose of using Elliptic Curve Cryptography (ECC) to control the access to data over sensor networks so that they presented their implementation of ECC in TelosB sensor network platform and evaluated their results by comparing it with the results of [18] and [19]; in [25] the paper is addressed to introduce the idea of SafeShare that consists of controlling the access by encapsulation of shared data so that their point of view consists of using the ABE to encrypt, encapsulate, audit and log the data in order to define a perform access control policy; other works go to the fact of using data content to control the access such as it is pointed out in [30] and [33].

3. PRELIMINARIES

Before going far in our work, we like to give you a complete grasp about some general concepts that we used in this report.

3.1. Standard deviation

It is a mathematical concept that gives the measure of dispersion of a specific population starting from its mean which can be regarded as the average of the population's values, however, the standard deviation is linearly related to the multiplication of the individuals over the population space; the more the individuals are spread the higher is the deflection; the following formula is the used one to calculate the standard deviation of such population of size n.

$$S= \sqrt{\frac{\sum_{i=0}^{n}(X_i - \bar{X})^2}{n-1}} \qquad (1)$$

Where the $\bar{X}$ presents the mean that is calculated using the following formula

$$\overline{X} = \frac{\sum_{i=0}^{n} X_i}{n} \qquad (2)$$

And X_i is an individual of the population.

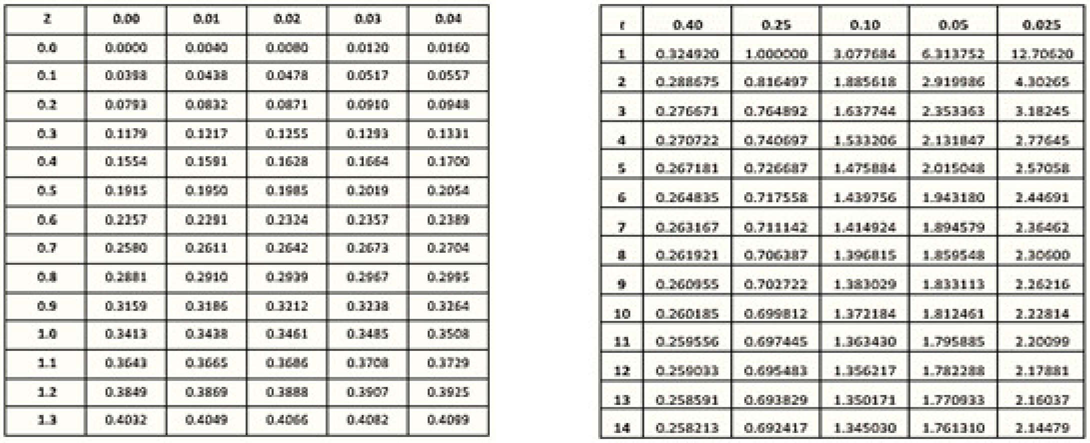
The standard deviation that is used in many cases such as in [16] where the authors proposed a new approach for selection of best threshold where the goal is to obtain better results for image segmentation and evaluated their results by comparing it with other conventional methods in term of several criterions such as the number of misclassified pixels; in [8] the authors proposed and evaluated a new query performance predictor for retrieval models using the standard deviation by testing several confidence levels; another use of standard deviation in information sciences is presented in [34] where the authors presented a standard deviation model to answer the problem of failure data in software reliability that presents a major problems in money costs and costumer satisfactions.

3.2. Confidence interval

It is an inferential statistical measurement that represents an interval of probability that such population individual can fall in basing on three essential parameters: the population's mean, the standard deviation and a specific percentage called confidence level. The confidence interval is calculated as follows:

$$CI= mean \pm marge\_error \quad (3)$$

Where the merge error presents the remainder between the mean and the extremities of the interval, the equation used to compute the merge error has two different cases: the case of a sample which has a size less than 30 and the one which accepts a size more than 30, the initial difference resides in special value called t-value in the first case and z-value in the second one, these two values are pulled from two different tables as shown the figure 1 bellow:



z	0.00	0.01	0.02	0.03	0.04
0.0	0.0000	0.0040	0.0080	0.0120	0.0160
0.1	0.0398	0.0438	0.0478	0.0517	0.0557
0.2	0.0793	0.0832	0.0871	0.0910	0.0948
0.3	0.1179	0.1217	0.1255	0.1293	0.1331
0.4	0.1554	0.1591	0.1628	0.1664	0.1700
0.5	0.1915	0.1950	0.1985	0.2019	0.2054
0.6	0.2257	0.2291	0.2324	0.2357	0.2389
0.7	0.2580	0.2611	0.2642	0.2673	0.2704
0.8	0.2881	0.2910	0.2939	0.2967	0.2995
0.9	0.3159	0.3186	0.3212	0.3238	0.3264
1.0	0.3413	0.3438	0.3461	0.3485	0.3508
1.1	0.3643	0.3665	0.3686	0.3708	0.3729
1.2	0.3849	0.3869	0.3888	0.3907	0.3925
1.3	0.4032	0.4049	0.4066	0.4082	0.4099

a) A sample from Z-table

r	0.40	0.25	0.10	0.05	0.025
1	0.324920	1.000000	3.077684	6.313752	12.70620
2	0.288675	0.816497	1.885618	2.919986	4.30265
3	0.276671	0.764892	1.637744	2.353363	3.18245
4	0.270722	0.740697	1.533206	2.131847	2.77645
5	0.267181	0.726687	1.475884	2.015048	2.57058
6	0.264835	0.717558	1.439756	1.943180	2.44691
7	0.263167	0.711142	1.414924	1.894579	2.36462
8	0.261921	0.706387	1.396815	1.859548	2.30600
9	0.260955	0.702722	1.383029	1.833113	2.26216
10	0.260185	0.699812	1.372184	1.812461	2.22814
11	0.259556	0.697445	1.363430	1.795885	2.20099
12	0.259033	0.695483	1.356217	1.782288	2.17881
13	0.258591	0.693829	1.350171	1.770933	2.16037
14	0.258213	0.692417	1.345030	1.761310	2.14479

b) A sample from T-table

Figure. 1. A samples from t-table and z-table

However, the extraction of values from these two tables is different, meanwhile, in our work we use z-table because of our population has 2 000 individuals in which are divided into 10 groups each one has more than 150 individuals, the computation of error marge passes by two steps:

- Extracting z-value from the table; for that, we must compute the α/2 value where the α is the confidence level, let's get an example of confidence level of 90%, the α/2 value is 0.90/2= 0.45, after that we search the closest value in the table, we find 0.4495 and 0.4505, then for each one of these values we calculate the corresponding row + the corresponding column and we get 1.64 and 1.65, finally the z-value equals to (1.64+1.65) /2= 1.645

- Now we have the z-value, the merge error is calculated using the following formula:

$$\text{Error\_marge= z-value x } \frac{S}{\sqrt{n}} \quad (4)$$

Where the S is the standard deviation and the n is the size of the sample

Another value could be derived from the standard deviation called the standard error that represents the distribution of the sample and it is figured using the formula 5 as follows:

Standard error = $\frac{S}{\sqrt{n}}$ (5)

3.3. Primitive root

In informatics security the primitive root is an important concept used in several cases, especially in the case of sharing the keys in public key cryptography schemes; formally a primitive origin of a number P is the number that satisfies the following attribute:

r is a primitive root of P => □ i, j ∈ ℕ, if i ≠j than ri mod P ≠ rj mod P

Nevertheless, in mathematics there is no accurate way to compute a primitive root of a number, instead, there is a method to verify if such number r is a primitive origin of a number P as shown the following code:

```
Procedure isPrimitiveRoot (number r; number P)
Begin
Compute ℓ (P);
Decompose ℓ (p) to a set of prime factors
For each prime factor fi do
            Compute mi= $r^{\ell (P)/fi}$ mod P
If all mi ≢ 1 mod P & mi ≢ -1 mod P then r is primitive root of P
End.
```

The most known algorithm which uses the primitive root is the famous Deffie-Helman algorithm for sharing secrete keys because of his special characteristic that is known as discrete logarithm problem where it is proved that for a number r being primitive root of a number P, if we know r, P, and the result of ra mod P we could never conclude the number a

3.4. Hierarchical identification

The big data knows comes with real evolution not only in term of data volume, but also in term of number of users which makes the identification of them a crucial problem and implies the search for new technics of choosing identities; one of these technics is a promising and new method called Hierarchical Identification that aims to benefit from different information concerning the users such as their groups so that the identities are depending on these information using the concatenation process as shows the figure 2 bellow:

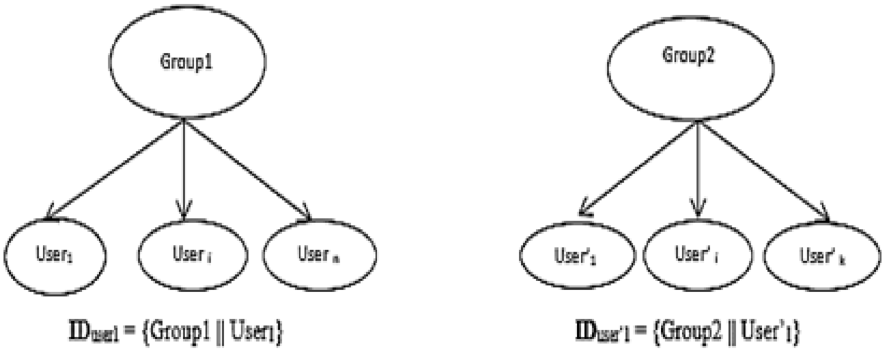

Group1
$User_1$
$User_i$
$User_n$
ID_{user1} = {Group1 || User1}
Group2
$User'_1$
$User'_i$
$User'_k$
$ID_{user'1}$ = {Group2 || User'1}

Figure. 2. Hierarchical Identification

This method has a major advantage resides in the ability of using the same identifier for multi users in different groups which allow the identification of big number of users with small size of identities and that can be useful in many cases that are related to the identification such as authentication mechanisms.

4. OUR APPROACH

Our advance is founded on three independent processes: defining access policy by computing the access control matrix and process of sharing the access rights.

4.1. Computing the access control matrix

This process is based, as the figure 3 shows above, on five steps: identification, normalization of identities, calculation of confidence interval for each group, calculation of digital signature for each user and ultimately determine the access rights by defining the matrix of access rights; in the remainder of this section we will detail each one of the stairs:

4.1.1. Identification

In this step, we target to get the identities of users utilizing the hierarchical identification mechanism in society to afford a standard configuration and size of the identities, we pass an address range of 10000 identities for each group using a concatenation operation between the group's ID where the user belongs and the genuine identity of the user, for example, let's consider a user with IDu= 0001 who belongs to a group which has the IDg= 01, the used identity of the user in our organization will be IDfinal= Edge || IDu= 010001= 10001.

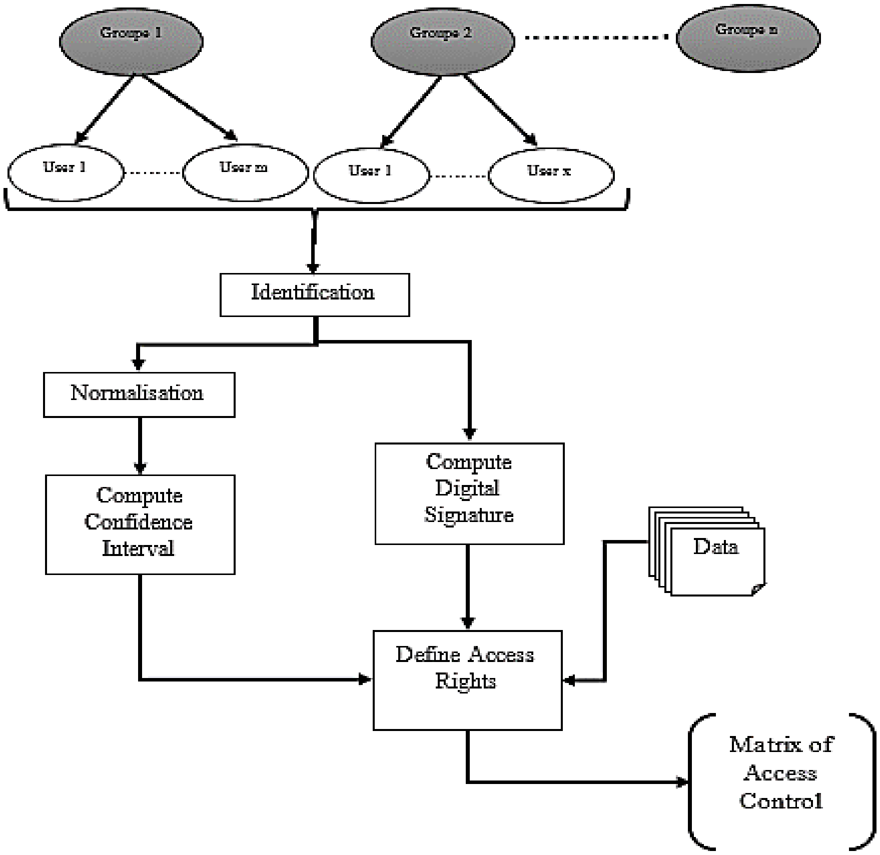


Figure. 3. Process of computing the access control matrix

4.1.2. Normalization

We can notice from the formula 4 that the standard deviation has a role in the process of computing the error marge, from its position in the formula, it's clearly shown that the error merge is linearly related to the standard deviation; otherwise, the more bigger is the standard deviation the more bigger is the error marge, by consequence, the more larger is the confidence interval, for that we suggest a normalisation of identities of the groups in order to create a less propagation rate in the range corresponded to each group so that instead of a maximum difference of 10000 between the extremities of a values of group, we diminish this value to 1 by dividing the identities by 10000. Getting hold of the example of a group in which the ID values go from 00001 to 10000, the normalized values go from 0.0001 to 1.0000; another normalization is used at the level of groups' sizes in order to preclude the influence of the great number of users in the group in the process.

4.1.3. Computing confidence interval

Our system defines for each group a specific confidence interval within the identity range of the group, this interval is estimated using various parameters, leading off from the standard deviation, and so setting the confidence layer, after that computing the merge error, finally utilizing the group means to get the final confidence level; all the computations in this step use the normalized values instead of the real ones.

4.1.4. Computing the digital signatures

This step is independent from the two precedent steps so that it can be executed in parallel with them, however, this step uses the concept of primitive root defined in the section 2.3 in such way that guarantees the unicity of signature for each user; to do that our system, firstly, generates for each group a big prime number P and find one of his primitive roots R, after that, for each user the generated signature equals to RIDfinal mod P; we choose this formula for two reasons: first, since R is primitive root of P then for two different users we will never get the same signature; and the second reason is that to protect the IDfinal of the user because of its major property known as Discrete Logarithm Problem cited in section 2.3, the IDfinal used in this step is in the real form and not the normalised one, thus, our system consists of generating the P as big prime because of two reasons too: first, there is no exact mathematical way to compute the primitive root but instead of that there is a way to verify if such number is primitive root as the procedure in section 2.3 shows using ℓ(P), so that we choose the P as prime to optimise the computation because ℓ(P) in this case equals to P-1; the second reason of choosing P as prime is also an optimisation reason because the researches proved that for P prime, we have a probability of 0.50 to generate a primitive root between 0 and ℓ(P).

4.1.5. Generation of access control matrix

This is the most important step of this process where the access rights are defined starting from the confidence interval of each group, the normalised identity of the user and his own signature; to do that our system conducts a set of tests for each group and each user by verifying if the normalised identity of a user belongs to the confidence interval of the group in order to know in which group the user belongs; once the system defines that, it start comparing the signatures of the data with the one of the user so that if are equals, the user will have full access and all the rights on the data that is considered as his own; else the user will have access on read only on the data that is considered in this case as shared data with him; for the other groups that the user doesn't belong, he will get no access right to their corresponding data; at the end of this process a matrix user x data is generated that resumes the access control policy.

The following code resume the process of creation of access control matrix by our system:

```
Algorithm AccessControlMatrix ( )
Input: $ID_g$ : groups' ID, $ID_u$: users' ID;
Output: M : matrix of access rights;
Begin
For each $group_i$ from IDg do
    For each $user_j$ from IDu that belongs to $group_i$ do
        $IDfinal_j \leftarrow group_i \; || \; user_j$;
        IDfinal←IDfinal+{$IDfinal_j$};
    End for;
End for;
For each $group_i$ from IDg do
    $sizegroup_i \leftarrow \frac{sizegroup_i}{2000}$
    $CI_i$ = Compute_confidence_interval ();
    Generate big prime number P and primitive root r
End for;
For each $user_j$ in IDfinal do
    $IDuser_j \leftarrow \frac{IDfinal_j}{10000}$
```

```
Compute $signature_j \leftarrow r^{IDfinal_j}$ mod P
End for;
M←new Matrix [number of users] [number of documents]
For each $group_i$ from IDg do
For each $user_j$ in IDfinal do
 If ($IDuser_j \in CI_i$) then
 For each $Document_k$ corresponds to $group_i$ do
  If ($signature_j = signature_k$) then
   M [j] [k] ← "read/write access";
  Else M [j] [k] ← "read only access";
  End if;
 End for;
 Else for each $Document_k$ corresponds to $group_i$ do
    M [j] [k] ← "no access right";
  End for;
 End if;
End for;
End for;
End.
```

4.2. Process of sharing the access rights

This procedure has been summed in order to answer some other problem that came to mind; what if such user decide to grant some other user to have write access right to his own data?, Otherwise, we aim by adding this process to allow for users of our system to share the same rights on same datum, to do that, our system uses Deffie-Helman algorithm of sharing cryptographic keys, but in our case to share the signatures between users; first of all, our system generates randomly a big prime number Q and a primitive root r then computed a primary signature for each of the users that will have the same access right S = $r^{ID_{user1}}$ mod Q that corresponds to user1, finally the system computes the final signature S_{final} = $S^{ID_{user2}}$ mod Q and signs the data with it. In this process, our system generates a new big prime Q and his primitive root and utilize them in the stead of the P that corresponds to the group where user1 belongs in order to protect the original signatures of the both users because it is used to sign other data that the users don't want to share rights with each other. The following process resumes this step:

```
Algorithm RightsSharing ()
Input $ID_{user1}$, $ID_{user2}$: users identities
Output $S_{final}$: final shared signature
Begin
Generate randomly a big prime Q and a number r < (Q-1);
While (r is not a primitive root of Q) do
Generate randomly a new r < (Q-1);
End while;
Compute S ← $r^{ID_{user1}}$ mod Q;
Compute $S_{final}$ ← $S^{ID_{user2}}$ mod Q;
Sign data with $S_{final}$;
End.
```

However, in this approach we choose to sign the data independently of its content, unlike the work presented in [30] because of two reasons: first, is to protect the privacy of data by perturbing the name in order to hide the real extension of the documents; and secondly, is to prevent problems of distrust like the famous one that Dropbox had recently because of its policy against violation of copyrights where a client claimed to the company from reading the content of his own data via Tweeter after the company prevent him from storing a document because of copyright violation[1].

5. EXPERIMENTS AND RESULTS

We take a set of experiments by building up a framework consists of 2000 users where each one has ten files stored in our system with a total of 20000 documents which gives an access control matrix of 2000 x 20000 that equals to 40 million right; the users are divided into 10 groups; this section is reserved for the introduction of a set of results using various parameters. But before going to the results, we will present the details of our dataset as shown in table 1

Table. 1. Dataset details used in our system

Group	Number of users	Range of identities	Range of normalised identities	Corresponding normalised Mean	Corresponding normalised standard deviation
Group 01	181	[0001…09999]	[0,0001…0,9999]	0.096	2.511
Group 02	234	[10000…19999]	[1,0000…1,9999]	1.100	3.377
Group 03	205	[20000…29999]	[2,0000…2,9999]	2.081	11.140
Group 04	209	[30000…39999]	[3,0000…3,9999]	3.020	23.819
Group 05	190	[40000…49999]	[4,0000…4,9999]	4.101	25.117
Group 06	184	[50000…59999]	[5,0000…5,9999]	5.098	25.649
Group 07	191	[60000…69999]	[6,0000…6,9999]	6.101	25.311
Group 08	221	[70000…79999]	[7,0000…7,9999]	7.096	23.672
Group 09	193	[80000…89999]	[8,0000…8,9999]	8.067	34.671
Group 10	193	[90000…99999]	[9,0000…9,9999]	9.098	34.774

As we notice in table 1, the mean is entirely related to the distribution of the values in their specific range and does not necessarily show the core of the range, and we notice also that the standard deviation is always out of the range of the sample because of the use of the ability of two during his computation.

5.1. Results

We carry on a set of comparisons organized in two steps: first, we confront a comparison between domains in order to study the influence of the distribution of the sample in our approach, secondly, we study the effect of choosing the confidence level in our approach, and finally, we evaluate our system by comparing it with other conventional work.

[1] http://assoquebecois.com/2014/04/01/dropbox-clarifie-sa-politique-sur-lexamen-des-dossiers-partages-pour-les-questions-dmca/

The next table shows the result of average of the access rate by domain in many experiments on normalized identities and real ones.

Table. 2. Results of comparison between groups in term of normalised and not normalised values and distribution of samples

Group	Standard error (distribution of sample)	Average of Access rate (%)	
		Normalized values	Not normalized values
Group 01	0.187	68.04	39.66
Group 02	0.220	77.48	89.00
Group 03	0.778	98.53	82.33
Group 04	1.647	97.12	114.33
Group 05	1.822	100.00	135.00
Group 06	1.891	100.00	152.66
Group 07	1.831	100.00	321.50
Group 08	1.592	100.00	341.00
Group 09	2.495	122.27	368.00
Group 10	2.503	119.42	364.00
Average	1.497	98.28	200.74

As the table 2 shows, the case of using normalised values gives better results than the one of not normalised values, however, we can see that the access policy exceed the limits in two groups: the Group 09 with an average of 122.27% (about 239 user could access to data instead of 193) and Group 10 with an average of 119.42% (230 users could access to data). Meanwhile the groups 05, 06, 07, and 08 give the best result with no error; meanwhile, the standard error is relatively related to the value of the standard deviation and that’s what influence the value of access rate. The more the standard error is big, the more the number of authorised users to access is big. As the results indicates that the normalised values’ case is widely better than the other case, we will in the rest of this section focus our experiments in only the normalised case.

In table 4 below, we will detail the results of the admission rate by group in term of the chosen confidence level in which we used many confidence levels and we take the ones that give an excited results in some of our groups, the following board indicates the chosen confidence levels with the corresponding z-value of each ace.

Table. 3. Corresponding z-value for each chosen confidence level

Confidence level	**20%**	**28%**	**28%<<29%**	**30%**	**31%**
z-value	0.255	0.355	0.365	0.375	0.385

Table. 4. Access rate by domain in term of chosen confidence level

Confidence level **Domain**	**20%**	**28%**	**28%<<29%**	**29%**	**30%**	**31%**
Group 01	48.62	68.51	70.72	71.82	72.36	76.24
Group 02	55.55	78.20	79.91	81.20	84.19	85.90
Group 03	98.54	98.54	98.54	98.54	98.54	98.54
Group 04	97.13	97.13	97.13	97.13	97.13	97.13
Group 05	100.00	100.00	100.00	100.00	100.00	100.00
Group 06	100.00	100.00	100.00	100.00	100.00	100.00
Group 07	100.00	100.00	100.00	100.00	100.00	100.00
Group 08	100.00	100.00	100.00	100.00	100.00	100.00
Group 09	99.48	111.39	122.80	140.41	165.80	193.78
Group 10	100.00	100.00	108.81	124.87	136.79	146.63

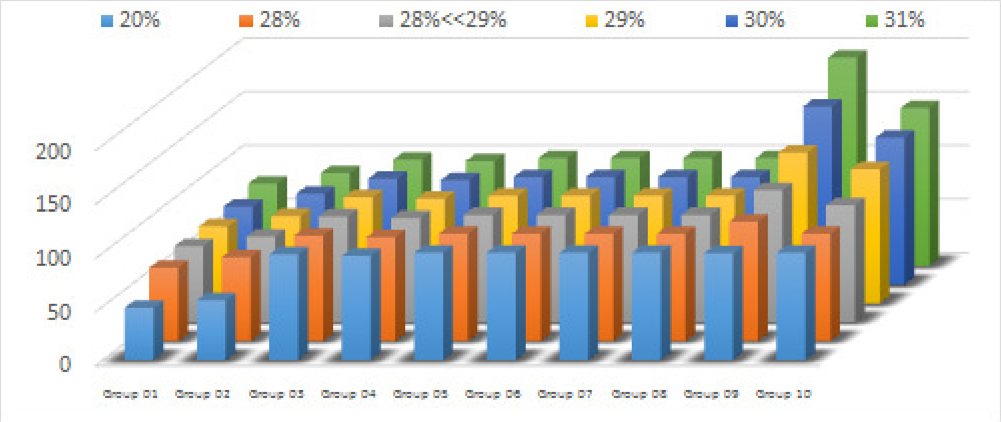


Figure. 4. Access rate by domain in term of chosen confidence level

As the table 4 and figure 4 show, each one of the confidence levels that we choose presents some good results in some groups and in the same time bad results in other groups; the best confidence level for groups group 01 and group 02 is 31% with rate of access of 76.24% in group 01 (138 user from 181) and 85.90% for group 02 (202 users from 234) while this level presents the worst results in programming group with 193.78% of access rate (374 users from 193 authorised), instead of that, the groups programming and security gives best results with less level of confidence using 20% of confidence level with 99.48% for programming (192 users from 193 authorised) and full access rate without error for security; meanwhile, the other groups such as data mining and natural sciences gives excited results without been influenced of the value of confidence level.

The following table presents the results of average of access rate and error rate between domains in term of variation of confidence level in normalised values where the positive value of error rate means that there is less users have access than the authorised ones and negative value means that there is more users that have access than the authorised ones, otherwise, the positive error rate means that there are users who must have access but our system doesn't allow to them to get access while the negative value means that there are some users must not access to data but our system allows to them to have access.

Table. 5. Results of average of access rate and error rate in term of confidence level variation

Confidence level (%)	Average of access rate (%)	Error rate (%)
20	89.92	10.08
28	95.37	4.63
28<<29	97.73	2.27
29	101.39	- 1.39
30	105.47	- 5.47
31	109.81	- 9.81

From the table 6 we can clearly notice that the confidence level between 28% and 29% gives better results with an average rate of access about 97.73% even if it represents some weaknesses in the last two domains where the access rate exceeds the 100 % (122.79% for 9 domain (44 unauthorized users), and 108.29% for 10 (16 unauthorized users)), the reason of why we didn't determine the exact value between 28 and 29 is that because all values within this range gives the same z-value which is about 0.365. The use of confidence level equals to 20% presents a major advantage because of all the values of access rate doesn't exceed the 100%, which means for all data there is no unauthorized access while it presents the worst result in term of error rate with more than 10% of authorized users could not access to data that must get access to because of the less access rate in the first domains where only 48.61% (only 88 users) of authorized users could access in 1 domain and 55.55% (only 130) could access in 2 domain. So, as the table shows, once we defined a confidence level starting from 29% the results became more badly (average of 28 of unauthorized users could access to data for 29%, 110 to 30%, and 197 to 31%).

After introducing a set of outcomes using a variation of parameters, we put our system in confrontation with a set of conventional works in the image of the system presented in [18] named TinyECC, and the one shown in [27] under the name ECC-AC in term of time of generating a signature and time of verifying the signature, however, our system generates a signature of average of size of 128 bits because of the role of a prime number of sizes of 2048 bits and primitive root of 1024 chips; the following table introduces the effects of time of generation and verification of signatures

Table. 6. Comparison of time of generation and verification of signatures

Approach	Time of generation of signature (s)	Time of verification of signature (s)
DSCI-AC	0.031	0.073
ECC-AC	3.35	6.78
TinyECC	8.24	16.28

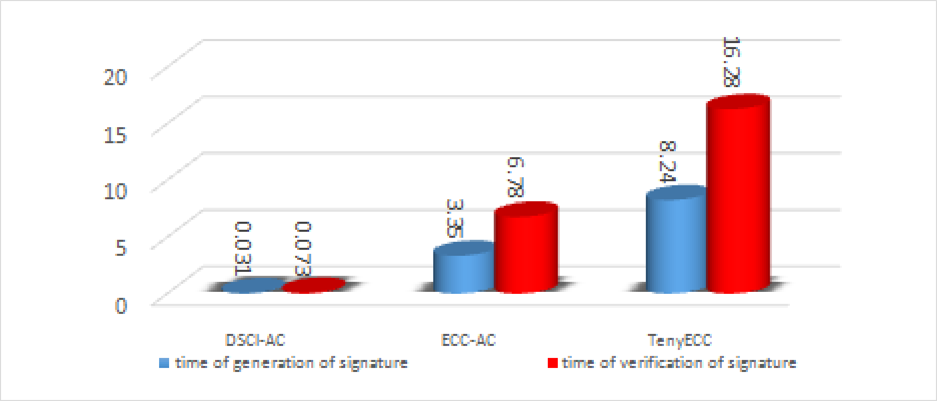


Figure. 5. Results of comparison of time of generation and verification of signatures

As the table 6 and figure 5 indicate, our system doesn't take much time for generating or verifying the signatures and that's due to the fact of using a simple computations to generate signatures and also the signature are used to perturb the data via their names which make its verification low costs because the system doesn't need to treat the data content to have signature, thus, in our approach only the server is the responsible of computing and verifying the signatures which leads us to eliminate the time of connection for users. In the other hand; we notice clearly that the time of generation of a signature is less than the one of its verification because the generation is based only on one and one only operation while the verification takes more time because of the number of operations resides on searching if the user belongs to the group in order to define if he has already the access or not then compare the two signatures to define the right that he has.

5.2. Limitations and weaknesses

As a security issue, our scheme could not present a safety sure state, nonetheless, our scheme has some restrictions that we could resume below:

- Confidence level: the use of a unique confidence level for all groups presents a limitation as shows the results above where the same confidence level prevent some authorized users from accessing their shared data in some groups and allows in the same time unauthorized users to access to data in other groups.

- Causing the server the only one who can generate and deploy signatures, making it easy to take on the role of man in the middle if we count the server as honest attacker, in fact, the server knows every signature used in the organization which allows for it to have broad access to all information, and that may be considered as assault of privacy requirements at some higher degrees of protection.

6. CONCLUSION

In this paper, we ushered in a new glide path of applying digital signature and the confidence interval in order to answer three essential questions: how could we control the approach to information that we don't hurt even the control on?, To answer it, we first divided the users into groups by their domains then compute for each group its own confidence interval that we used in our system in parliamentary procedure to ascertain who has access to data and who doesn't, after determining which user has access to the data, another question came to mind; for the users who have access to data, which right should they have, is that full access or read only access? We answered this question by using the digital signature generated using another mathematical concept called primitive root basing on prime numbers and random theories in order to precisely which access right each user must take; then by assisting these two questions we could define the last access control matrix; the final question that we answered in this study is that if such user decide to afford full access on his data for another user, how could we ensure that? To respond that we offered the use of Deffie-Hellman algorithm of sharing cryptographic keys in order to permit users to partake in the same signature by consequence have the same access right on the same data.

As future work, we will usher in new models of using meta-heuristics technics to improve the results of this work by searching for the appropriate assurance level for each group we also will give other models using cryptography whose purpose is to prevent the server from recognizing the genuine signatures of the users. In the final stage, it only remains to mention that the security in Big Data is all grounded on trust then that no trust no security.

REFERENCES

[1] Arunkumar, S., Raghavendra, A., Weerasinghe, D., Patel, D., & Rajarajan, M. (2010, October). Policy extension for data access control. In Secure Network Protocols (NPSec), 2010 6th IEEE Workshop on (pp. 55-60). IEEE.

[2] Astorga, J., Jacob, E., Huarte, M., & Higuero, M. (2012). Ladon 1: end-to-end authorisation support for resource-deprived environments. Information Security, IET, 6(2), 93-101.

[3] Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors.Bmj, 331(7521), 903.

[4] Bagheri, E., Babaei, S., & Khayyambashi, M. R. (2009, August). A new method for consistency of access control in web services. In Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on (pp. 567-569). IEEE.

[5] Camenisch, J., Mödersheim, S., Neven, G., Preiss, F. S., & Sommer, D. (2010, June). A card requirements language enabling privacy-preserving access control. In Proceedings of the 15th ACM symposium on Access control models and technologies (pp. 119-128). ACM.

[6] Chen, Y. R., Chu, C. K., Tzeng, W. G., & Zhou, J. (2013, January). Cloudhka: A cryptographic approach for hierarchical access control in cloud computing. InApplied Cryptography and Network Security (pp. 37-52). Springer Berlin Heidelberg.

[7] Crampton, J. (2009). Cryptographically-enforced hierarchical access control with multiple keys. The Journal of Logic and Algebraic Programming, 78(8), 690-700.

[8] Cummins, R., Jose, J., & O'Riordan, C. (2011, July). Improved query performance prediction using standard deviation. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 1089-1090). ACM.

[9] Curé, O., Kerdjoudj, F., Le Duc, C., Lamolle, M., & Faye, D. (2012, September). On the potential integration of an ontology-based data access approach in NoSQL stores. In Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on (pp. 166-173). IEEE.

[10] Di Vimercati, S. D. C., Foresti, S., Jajodia, S., Paraboschi, S., & Samarati, P. (2007, September). Over-encryption: management of access control evolution on outsourced data. In Proceedings of the 33rd international conference on Very large data bases (pp. 123-134). VLDB endowment.

[11] Guo, J., Baugh, J. P., & Wang, S. (2007). A group signature based secure and privacy-preserving vehicular communication framework. Mobile Networking for Vehicular Environments, 2007, 103-108.

[12] Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006, October). Attribute-based encryption for fine-grained access control of encrypted data. InProceedings of the 13th ACM conference on Computer and communications security (pp. 89-98). ACM.

[13] Keathley, E. F. (2014). Big Data and Bigger Control Issues. In Digital Asset Management (pp. 99-115). Apress.

[14] Kelani Bandara, K. B. P. L. M., Wikramanayake, G. N., & Goonethillake, J. S. (2007, August). Optimal selection of failure data for reliability estimation based on a standard deviation method. In Industrial and Information Systems, 2007. ICIIS 2007. International Conference on (pp. 245-248). IEEE.

[15] Khalil, I., Khreishah, A., & Azeem, M. (2014). Consolidated Identity Management System for secure mobile cloud computing. Computer Networks,65, 99-110.

[16] Li, Z., Cheng, Y., Liu, C., & Zhao, C. (2010, March). Minimum Standard Deviation Difference-Based Thresholding. In Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on (Vol. 2, pp. 664-667). IEEE.

[17] Li, M., Yu, S., Zheng, Y., Ren, K., & Lou, W. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. Parallel and Distributed Systems, IEEE Transactions on, 24(1), 131-143.

[18] Liu, A., & Ning, P. (2008, April). TinyECC: A configurable library for elliptic curve cryptography in wireless sensor networks. In Information Processing in Sensor Networks, 2008. IPSN'08. International Conference on (pp. 245-256). IEEE.

[19] Malan, D. J., Welsh, M., & Smith, M. D. (2004, October). A public-key infrastructure for key distribution in TinyOS based on elliptic curve cryptography. In Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004. 2004 First Annual IEEE Communications Society Conference on (pp. 71-80). IEEE.

[20] Miklau, G., & Suciu, D. (2003, September). Controlling access to published data using cryptography. In Proceedings of the 29th international conference on Very large data bases-Volume 29 (pp. 898-909). VLDB Endowment.

[21] Ortiz, P., Lazaro, O., Uriarte, M., & Carnerero, M. (2013, June). Enhanced multi-domain access control for secure mobile collaboration through Linked Data cloud in manufacturing. In World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a (pp. 1-9). IEEE.

[22] Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2014). Sensing as a service model for smart cities supported by internet of things. Transactions on Emerging Telecommunications Technologies, 25(1), 81-93.

[23] Shtok, A., Kurland, O., & Carmel, D. (2009). Predicting query performance by query-drift estimation. In Advances in Information Retrieval Theory (pp. 305-312). Springer Berlin Heidelberg.

[24] Stevens, G., & Wulf, V. (2009). Computer-supported access control. ACM Transactions on Computer-Human Interaction (TOCHI), 16(3), 12.

[25] Thilakanathan, D., Calvo, R., Chen, S., & Nepal, S. (2013, December). Secure and Controlled Sharing of Data in Distributed Computing. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on(pp. 825-832). IEEE.

[26] Tu, S. S., Niu, S. Z., Li, H., Xiao-ming, Y., & Li, M. J. (2012, May). Fine-grained access control and revocation for sharing data on clouds. In Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International (pp. 2146-2155). IEEE.

[27] Wang, H., Sheng, B., & Li, Q. (2006). Elliptic curve cryptography-based access control in sensor networks. International Journal of Security and Networks, 1(3), 127-137.

[28] Wang, Z. H., Zhi, S. S., & Liu, H. M. (2012, July). MSHS: The mean-standard deviation curve matching algorithm in HSV space. In Machine Learning and Cybernetics (ICMLC), 2012 International Conference on (Vol. 3, pp. 1064-1069). IEEE.Yang, Y., & Zhang, Y. (2011, September). A generic scheme for secure data sharing in cloud. In Parallel Processing Workshops (ICPPW), 2011 40th International Conference on (pp. 145-153). IEEE.

[29] Yu, S., Wang, C., Ren, K., & Lou, W. (2010, March). Achieving secure, scalable, and fine-grained data access control in cloud computing. InINFOCOM, 2010 Proceedings IEEE (pp. 1-9). Ieee.

[30] Zeng, W., Yang, Y., & Luo, B. (2013, October). Access control for big data using data content. In Big Data, 2013 IEEE International Conference on (pp. 45-47). IEEE.

[31] Zhang, X., Liu, C., Nepal, S., Dou, W., & Chen, J. (2012, November). Privacy-Preserving Layer over MapReduce on Cloud. In Cloud and Green Computing (CGC), 2012 Second International Conference on (pp. 304-310). IEEE.

[32] Yang, Y., & Zhang, Y. (2011, September). A generic scheme for secure data sharing in cloud. In Parallel Processing Workshops (ICPPW), 2011 40th International Conference on (pp. 145-153). IEEE.

[33] Nabeel, M., Bertino, E., Kantarcioglu, M., & Thuraisingham, B. (2011, October). Towards privacy preserving access control in the cloud. InCollaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2011 7th International Conference on (pp. 172-180). IEEE.

[34] Bandara, K., Wikramanayake, G. N., & Goonethillake, J. S. (2007, August). Optimal selection of failure data for reliability estimation based on a standard deviation method. In Industrial and Information Systems, 2007. ICIIS 2007. International Conference on (pp. 245-248). IEEE.

EFFICIENT DISPATCHING RULES BASED ON DATA MINING FOR THE SINGLE MACHINE SCHEDULING PROBLEM

Mohamed Habib Zahmani[1], Baghdad Atmani[2] and Abdelghani Bekrar[3]

[1]Laboratory of Pure and Applied Mathematics,
University of Mostaganem, Mostaganem, Algeria
habib.zahmani@univ-mosta.dz
[2]Laboratoire d'Informatique d'Oran,
University of Oran 1, Oran, Algeria
atmani.baghdad@univ-oran.dz
[3]Laboratory of Industrial and Human Automation control,
Mechanical engineering and Computer Science,
University of Valenciennes and Hainaut-Cambresis, Valenciennes, France
abdelghani.bekrar@univ-valenciennes.fr

ABSTRACT

In manufacturing the solutions found for scheduling problems and the human expert's experience are very important. They can be transformed using Artificial Intelligence techniques into knowledge and this knowledge could be used to solve new scheduling problems. In this paper we use Decision Trees for the generation of new Dispatching Rules for a Single Machine shop solved using a Genetic Algorithm. Two heuristics are proposed to use the new Dispatching Rules and a comparative study with other Dispatching Rules from the literature is presented.

KEYWORDS

Data Mining, Decision Trees, Dispatching Rules, Single Machine, Scheduling, Genetic Algorithm

1. INTRODUCTION

In today's business environment, competition has become very fierce and the customers have become very demanding in terms of quality, cost and time. In this context of an increasingly globalized world, and in order to guarantee the survival of the enterprise it is necessary to enhance the manufacturing process. By doing so, the company secures a place in this highly-competitive environment. The manufacturing process goes hand in hand with the scheduling problem. The scheduling problem is NP-Hard due to the exponential number of solutions [1].

In this paper we focus on the Single Machine (SM) Scheduling Problem. Many methods have been proposed to solve this problem including its many variations, namely, static and dynamic,

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 199–208, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51516

with or without perturbations and a large set of objective functions (sum of tardiness, length of schedule, etc.) [1]–[4]. Still no proposed approach can solve it for all its variations.

One of the most popular methods used to solve scheduling problem are the Dispatching Rules (DR). These techniques are widely used [5]–[7] owing to their efficiency and their ability to quickly define a priority for each job waiting on a machine queue. This reactivity increases the systems responsivity and its fault tolerance in case of dynamic or perturbation scenarios.

An evolution of Dispatching Rules introduced in literature is Data Mining. DM is used in order to create new rules using previous problems and experiments. In this way, the previously acquired knowledge is transformed to new DR to be used for new scheduling problems. Among these approaches we notice some in particular Koonce [8], Shahzad [9] or Aissani [10]. In all these works DM is used and more often Decision Trees for the extraction of Dispatching Rules to be used to schedule jobs in a new problem or reschedule in case of perturbation. In these papers authors focus mainly on multiple-machines shops such as Flow Shop or Job Shop for the generation of new DR.

In this paper Genetic Algorithms (GA) are employed in pretreatment process by solving the Single Machine problem, and Decision Trees are called upon to extract hidden knowledge in the form of Dispatching Rules. Also are present two heuristics for the setup and use of the new Dispatching Rules since they are different from classic ones. Finally, a comparative study with other well-known DR is performed.

2. STATE OF THE ART

In planning and scheduling, Aytug & al. [11] distinguish two ways of using Data Mining. The first for decision support, where DM helps to identify the best DR since no one rule outperforms all others such as in the paper of Metan & al. [12]. In this approach the state of the system is continuously monitored and the Decision Support System changes the Dispatching Rules if need be to optimize the objective function.

As for the second way, DM is applied to face perturbations scenarios for example a machine breakdown or new jobs arrival. A recent study of Said & al. [13] where the DR is dynamically changed in order to minimize the impact of the perturbation.

A third trend initially introduced by Li [14] where Decision Trees are employed to generate new Dispatching Rules capable of mimicking a metaheuristic or even an exact method for the resolution of a scheduling problem. This approach is proposed for a Single Machine problem using DR. Authors use the LPT rule (Longest Processing Time) where the job having the longest processing time is processed first. Then DT algorithm is applied for the generation of the new DR. One drawback in this paper is the use of the LPT rule since it is a heuristic it is not capable of finding the best solution. Therefore as an alternative, we propose the use of a Genetic Algorithm for the problems resolution.

Other works based on the same idea use different solving methods and for other scheduling problems. For instance in [9], a Job Shop problem is addressed using Tabu Search. The TS algorithm is used at first to find a feasible solution, afterwards a data pretreatment of the solution is done. Finally, Decision Trees, based on the pretreated data, generate a new DR.

Balasundaram & al. [15] perform also a DR generation using DT in a Flow Shop environment. At first a simple heuristic is used to solve the problem comprising 5 jobs and 2 machines. For the scheduling, the makespan (end date of all jobs) is estimated by scheduling a job i before job j and vice versa. The combination minimizing the makespan is used. Then, DT are applied on the scheduling solution to find a new DR. Another idea of Khademi Zare & al. [16] using a hybrid algorithm combining Genetic Algorithm, Data Mining, fuzzy sets, similarity algorithm and attribute-driven deduction algorithm. DM is used to extract rules to help GA to boost its speed to reach the optimal solution.

After a thorough analysis of these approaches, two problems arise. The first resides in the fact that the new Dispatching Rule in a "if-then" form is only compared with the solving methods used in the first step, a heuristic in [15], Tabu Search [9] or Shortest Processing Time for Li & al. [14]. This comparison shows that the new DR performs nearly as efficiently as the heuristic/metaheuristic used for the resolution of the problem, with a certain degree of error. This difference is due to the bad decisions taken by the decision (inverse some jobs). Also, the problem of jobs order inversion is not addressed in any of the quoted approaches, which may have heavy consequences on the system's performance. To illustrate the problem we propose the following example.

Suppose that the new obtained DR is constructed using 3 jobs, J_1, J_2 and J_3. The sequence returned using a solving method is for example: J_2, J_3 and finally J_1. The new DR, contrary to a classic DR, can only compare jobs one by one (see Figure 1):

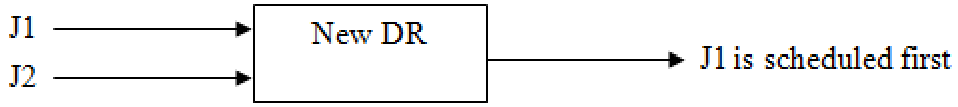


Figure 1. Defining sequence using the new DR

For example, while comparing the jobs the new DR will return the following results:

J_2 vs J_3: J_2 is scheduled first (correct decision)

J_3 vs J_1: J_3 is scheduled first (correct decision)

J_2 vs J_1: J_1 is scheduled first (wrong decision)

In case of a wrong decision it becomes impossible to construct a scheduling. According to the two first decisions, job 1 should be processed last. But with the third decision, being incorrect, there is an ambiguity. Should job 1 be processed last taking into consideration only the two first decisions, or must it be processed first, ignoring the two first decisions and taking only the third one into consideration.

A second problem is that no details are provided as to how to use the new DR in particular the approaches of Li and Shahzad [9], [14]. In those two papers the generated decision tree (DR) have a small size (2 nodes) because of the number of jobs used (5 jobs). But in a real complex problem the jobs number is much higher and consequently the tree size will be larger and more complex. Also the authors never compare the new DR with other rules from literature in order to evaluate its performance.

In this paper we propose two heuristics to take into consideration the bad decision problem and also a comparative study with some Dispatching Rules from the literature is performed.

3. Proposed Approach

To generate Dispatching Rules based on Decision Trees it is necessary to use an exact or approximate method to solve the problem. In this paper we use a Genetic Algorithm to solve the Single Machine Scheduling Problem, the solution is used by the Decision Tree to create a DR. The choice of GA is justified by its ability to quickly explore the research space, its proven results and wide use for such a problem [1], [17]–[19]. And as done by Li & Shahzad [9], [14] Decision Trees will be used for the DR generation.

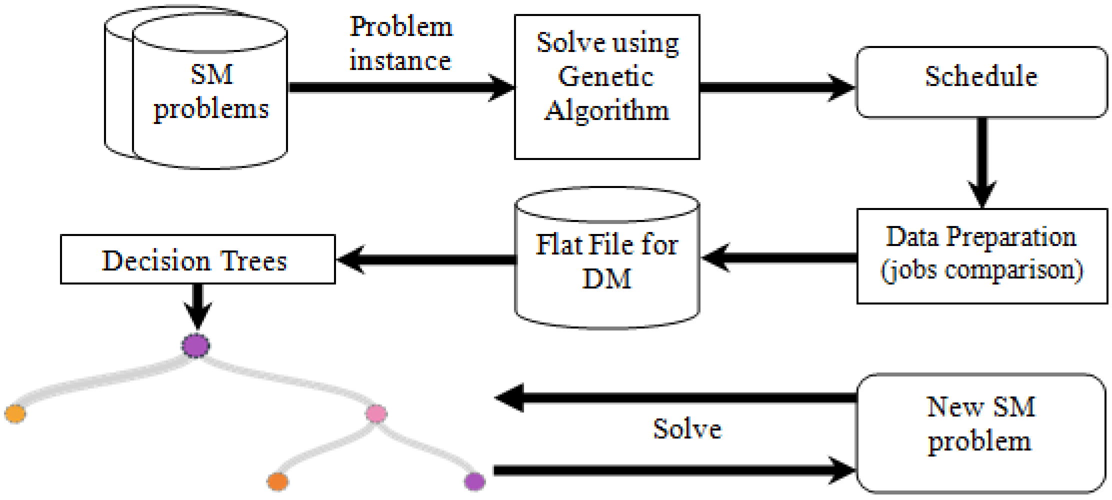


Figure 2. Proposed approach

3.1. Single Machine

The Single Machine Scheduling Problem with Total Weighted Tardiness (SMTWT) can be defined as follows. A set of n jobs $J = \{J_1, J_2, \dots, J_n\}$ to be processed on a machine. Each job J_i consists of a single operation having a processing time $PT_i > 0$, $i = 1, \dots, n$. The importance of a job is expressed using a positive weight $w_i > 0$, $i = 1, \dots, n$. The machine can process only one job at a time and a job's execution cannot be suspended. Each job is supposed to be finished before its deadline d_i. If not a penalty "tardiness" is then calculated where tardiness $T_i = \max\{S_i + PT_i - d_i; 0\}$. S_i denotes the J_i start time. The scheduling objective is to minimize the Total Weighted Tardiness:

$$\min TWT = \sum_{i=1}^{n} w_i T_i$$

3.2. Genetic Algorithm

Genetic Algorithms are used to explore a solution space by mimicking the biological process. They have been successfully applied in literature for several problems including Single Machine problem [18], [19]. The main components of GA are as follows:

- Solution encoding: a representation of solutions
- Initial population: generation of an initial population
- Fitness: measurement function for a given solution, total weighted tardiness in this case
- Selection: selection process for chromosomes to generate a new population
- Genetic operators: a genetic operator such as crossover and mutation are applied on the selected chromosomes in order to create new ones
- Replacement: natural selection of the members of population who will survive

In this paper, the implemented GA is mainly inspired from the one proposed by Armentano [18] with a modified initial population generation process. We suggest that the reader consult the paper for more details.

3.3. Decision Tree

For the generation of Dispatching Rules, we use the method of Li [14] which is adapted to our problem since the used attributes are not the same. Release time, start time, processing time and completion time in the case of [14]; and processing time, weight and due date in this paper. To explain the process we propose the following example:

Suppose there is 4 jobs, and the sequence returned by the GA is |3|0|2|1|

Table 1. Jobs attributes

Job N°	Processing Time	Weight	Due Date
0	15	3	80
1	20	2	30
2	35	6	50
3	7	8	60

A comparative table is constructed (see Table 2) based on Table 1 where jobs are compared one to another. So, job J_1 is compared with $N-1$, job J_2 with $N-2$ and so on. The size of the new table is equal to $\sum_{j=1}^{N-1}(n-j)$. This new table will be used as an entry by the Decision Tree to generate a new Dispatching Rule.

Table 2. Jobs comparison

Job1 N°	PT1	w1	d1	Job2 N°	PT2	w2	d2	Job1 1er
0	15	3	80	1	20	2	30	Yes
0	15	3	80	2	35	6	50	Yes
0	15	3	80	3	7	8	60	No
1	20	2	30	2	35	6	50	No
1	20	2	30	3	7	8	60	No
2	35	6	50	3	7	8	60	No

The C4.5 Decision Tree algorithm is the applied on the data of Table 2 to create the new DR. The new rules is an if-then form as follows:

$$\boldsymbol{if}\ w_2 \leq 6\ \boldsymbol{then}\ J_1\ is\ processed\ first$$

$$\boldsymbol{if}\ w_2 > 6\ \boldsymbol{then}\ J_2\ is\ processed\ first$$

3.4. New Dispatching Rule

In order to apply the new DR (DT), we propose two heuristics in order to take into consideration the bad decisions.

Proposed Heuristic 1 sort by number of accumulated of "yes"

```
Initialize vector NumberYes[N] to 0
For each Job i to N
    For each Job j to N
        Decision = JobiScheduledFirst(i, j) ;
        If Decision = Yes
            NumberYes [i] = NumberYes [i] + 1 ;
        End if
    End for
End for
Sort NumberYes in decreasing order
```

Proposed Heuristic 2 quick sort

```
While Iteration <= IterationsNumber
For each Job i to N
    For each Job j to N
        Decision = JobiScheduledFirst(i, j) ;
        If Decision = Yes
            Swap(i, j);
        End if
    End for
End for
End while
```

4. EXPERIMENTS AND RESULTS

In order evaluate the performances of the GA and the new DR, we focus on the Weighted Tardiness benchmark available in OR-Library (see http://people.brunel.ac.uk/~mastjjb/jeb/orlib/ wtinfo.html for more details). Where for each problem the best Total Weighted Tardiness is known allowing us by the same way to compare the two heuristics and other DR from literature. Experiments are performed on 125 problems with 40 jobs.

The population size of the GA I is set to 100 chromosomes and the size of the selected population for crossover/mutation is $I/2$ i.e. 50 chromosomes. While the number of iterations is set to 150 and if the best known solution is reached before the GA stops. The modified GA finds the best known solution for 32 problems, that is 25.6%.

We also conducted a comparison between the GA results and the best results. To do so, a gap using the following formula is calculated (1):

$$Average\ Difference = \left[\sum_{i=1}^{Problems\ Number} \frac{GA(i) - ORLibrary(i)}{ORLibrayr(i)}\right] / ProblemsNumber \quad (1)$$

$GA(i)$ is the score found using the Genetic Algorithm for a problem i.

$ORLibrary(i)$ is the best known score in OR-Library for a problem i.

It is worth mentioning that for some problems the best known score in the OR-Library is equal to 0. Thus it becomes impossible to calculate the gap, consequently, those problems are ignored. This reduces the number of problems to 107 (instead of 125 initially). On these 107 problems GA has 9.76% average difference with the OR-Library.

Once the problems are solved using GA, Decision Trees are applied to generate a new Dispatching Rule as explained in the proposed approach section. All the data of all the problems is gathered in one file for the learning process. Then, in order to apply the new DR the two heuristics PH1 and PH2 are used and a comparison is performed using the following formula (2) (see Table 3). The best heuristic in then applied for the comparison with literature Dispatching Rules.

$$Average\ Difference = \left[\sum_{i=1}^{Problems\ Number} \frac{H(i) - ORLibrary(i)}{ORLibrayr(i)}\right] / ProblemsNumber \quad (2)$$

$H(i)$ is a score found using a heuristic H for a problem i.

Table 3. Comparison of PH1 and PH2

PH1		PH2	
Aver. Gap (%)	NP	Aver. Gap (%)	NP
912	1	115,32*	106*

NP is the number of times where the heuristic finds the best results.

From these results we conclude that the second heuristic PH2 is much better than the first one. So, for the following experiments only this one is used for the comparison with the other DR.

The Dispatching Rules used for the comparison (see Table 4) study are:

- Shortest Processing Time (SPT)
- Longest Processing Time (LPT)
- Weighted Shortest Processing Time (WSPT), [2] this rule is best for the SMTWT problem.
- Earliest Due Date (EDD)
- Critical Ratio (CR)
- Mixed Dispatching Rule (MDR) propose dans [20]

Table 4: Comparison of PH2 with other DR

	PH2	**SPT**	**LPT**	**WSPT**	**EDD**	**CR**	**MDR**
Gap (%)	**115.32***	1203.43	4929.03	681.08	162.58	1828.13	162.58
NP	**81***	0	0	16	13	0	13

Based on this results, it is clear that the new Dispatching Rule returns better results in terms of average gap compared to the best known results of the OR-Library. Also in regards to time where it finds the smallest Total Weighted Tardiness value.

In Figure 3 we compare the results of PH2 and MDR being the best two DR. When the objective value is equal tp 0 it means that the optimal solution is reached by one of the two heuristics, if not the difference in terms of Total Weighted Tardiness is shown.

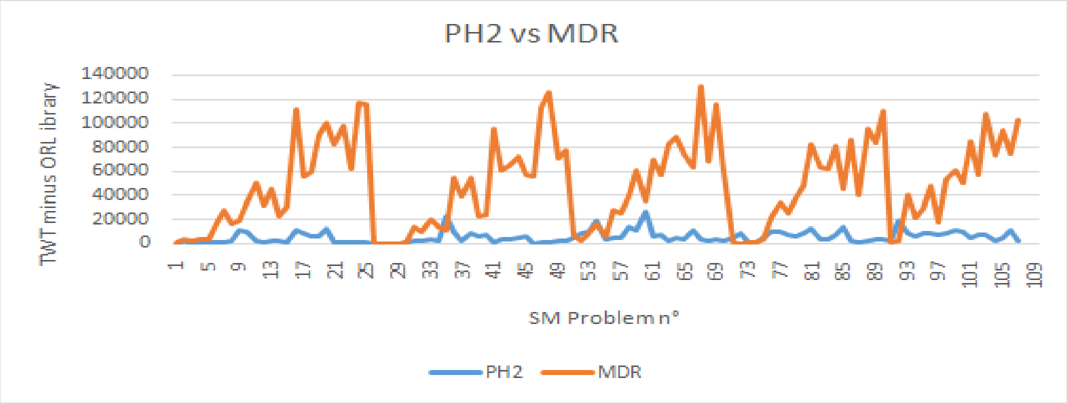


Figure 3. PH2 vs MDR

Also and in order to accurately measure the new DR the set of 107 problems is split in two parts 92/15. The 92 problems will be used for the learning process to create a new DR and other 15 for tests. The aim is to evaluate PH2 for new scheduling problems, results are shown in Table 5.

Table 5: PH2 vs classic DR for new problems

	PH2	SPT	LPT	WSPT	EDD	CR	MDR
Gap (%)	**63.59***	126.11	360.30	68.07	123.86	274.85	123.86
NP	**7***	0	0	**7***	1	0	1

In case of an entirely new Single Machine Scheduling Problem, the proposed heuristic PH2 has an average difference of 63.59% compared to the best known results. In terms of number of times where the best score is reached it performs as well as the WSPT rule. In Figure 4 we compare the results of PH2 and WSPT as done in Figure 3.

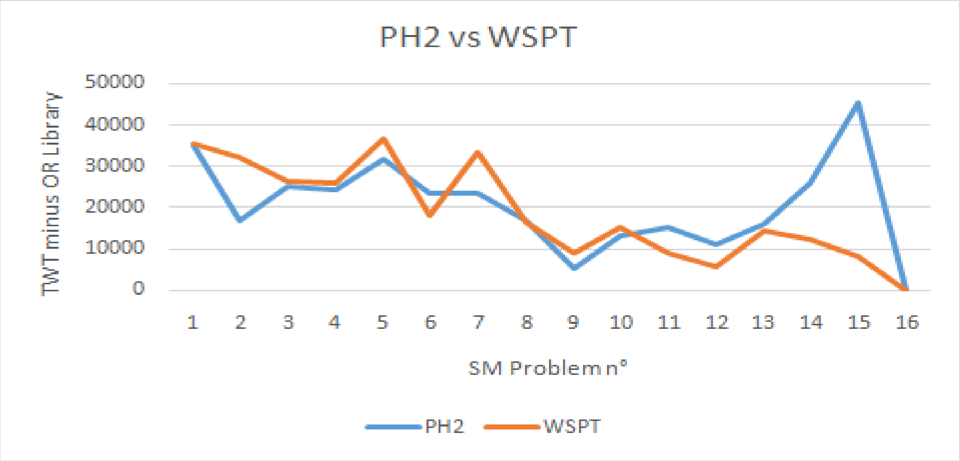


Figure 4. PH2 vs WSPT

From this experiments we prove the superiority of the proposed heuristic for the use of the Decision Tree as Dispatching Rule, while taking into consideration bad decisions. The proposed heuristic also outperforms WMDD [21], H2 et H3 [22] for the same set of data.

5. CONCLUSIONS

In this paper two heuristics for the use of Decision Trees as Dispatching Rule based on a Genetic Algorithm are proposed. The approach was tested for Single Machine problem with Total Weighted Tardiness objective. Experiments show the superiority of the proposed approach compared to some well-known DR for problems used in learning or completely new ones.

In perspective, an improvement of the proposed heuristic is possible. Also, knowing that OR-Library includes SM problems with 50 and 100 jobs, it is interesting to test the new heuristic for such problems. Finally, it might be interesting to consider more complex problems with multiple machines such as the Job Shop Scheduling Problem.

REFERENCES

[1] N. Liu, M. Abdelrahman, and S. Ramaswamy, “A Genetic Algorithm for Single Machine Total Weighted Tardiness Scheduling Problem,” Int. J. Intell. Control Syst., vol. 10, no. 3, pp. 218–225, 2005.

[2] E. J. Anderson and C. N. Potts, “Online scheduling of a single machine to minimize total weighted completion time,” Math. Oper. Res., vol. 29, no. 3, pp. 686–697, 2004.

[3] H. Zhu and H. Zhou, “Predictive Scheduling for a Single Machine with Random Machine Breakdowns,” in LISS 2013, Springer, 2015, pp. 753–758.

[4] J. Tian, R. Fu, and J. Yuan, “A best on-line algorithm for single machine scheduling with small delivery times,” Theor. Comput. Sci., vol. 393, no. 1–3, pp. 287–293, 2008.

[5] S. Nguyen, M. Zhang, M. Johnston, and K. C. Tan, “A Computational Study of Representations in Genetic Programming to Evolve Dispatching Rules for the Job Shop Scheduling Problem,” IEEE Trans. Evol. Comput., vol. 17, no. 5, pp. 621–639, Oct. 2013.

[6] T. Hildebrandt, D. Goswami, and M. Freitag, “Large-scale simulation-based optimization of semiconductor dispatching rules,” in WSC ’14 Proceedings of the 2014 Winter Simulation Conference, 2014, pp. 2580–2590.

[7] T. Hildebrandt, J. Heger, and B. Scholz-Reiter, “Towards improved dispatching rules for complex shop floor scenarios: a genetic programming approach,” in GECCO ’10: Proceedings of the 12th annual conference on Genetic and evolutionary computation, 2010, pp. 257–264.

[8] D. Koonce and S. Tsai, “Using data mining to find patterns in genetic algorithm solutions to a job shop schedule,” Comput. Ind. Eng., vol. 38, no. 3, pp. 361–374, Oct. 2000.

[9] A. Shahzad and N. Mebarki, “Discovering Dispatching Rules for Job Shop,” in 8th International Conference of Modeling and Simulation - MOSIM’10, 2010.

[10] N. Aissani, B. Atmani, D. Trentesaux, and B. Beldjilali, “Extraction of Priority Rules for Boolean Induction in Distributed Manufacturing Control,” Serv. Orientat. Holonic Multi-Agent Manuf. Robot. Stud. Comput. Intell., vol. 544, pp. 127–143, 2014.

[11] H. Aytug, S. Bhattacharyya, G. J. Koehler, and J. L. Snowdon, “A review of machine learning in scheduling,” IEEE Trans. Eng. Manag., vol. 41, no. 2, pp. 165–171, May 1994.

[12] G. Metan, I. Sabuncuoglu, and H. Pierreval, “Real time selection of scheduling rules and knowledge extraction via dynamically controlled data mining,” Int. J. Prod. Res., vol. 48, no. 23, pp. 6909–6938, Dec. 2010.

[13] N. Said, W. Mouelhi, and K. Ghedira, “Classification Rules for the Job Shop Scheduling Problem with Machine Breakdowns,” Int. J. Inf. Electron. Eng., vol. 5, no. 4, p. 300, 2015.

[14] X. Li and S. Olafsson, “Discovering Dispatching Rules Using Data Mining,” J. Sched., vol. 8, no. 6, pp. 515–527, Dec. 2005.

[15] R. Balasundaram, N. Baskar, and R. S. Sankar, “A New Approach to Generate Dispatching Rules for Two Machine Flow Shop Scheduling Using Data Mining,” Procedia Eng., vol. 38, pp. 238–245, 2012.

[16] H. Khademi Zare and M. B. Fakhrzad, “Solving flexible flow-shop problem with a hybrid genetic algorithm and data mining: A fuzzy approach,” Expert Syst. Appl., vol. 38, no. 6, pp. 7609–7615, Jun. 2011.

[17] F. Der Chou, P. C. Chang, and H. M. Wang, “A hybrid genetic algorithm to minimize makespan for the single batch machine dynamic scheduling problem,” Int. J. Adv. Manuf. Technol., vol. 31, no. 3–4, pp. 350–359, 2006.

[18] V. a. Armentano and R. Mazzini, “A genetic algorithm for scheduling on a single machine with set-up times and due dates,” Prod. Plan. Control, vol. 11, no. 7, pp. 713–720, Jan. 2000.

[19] M. Sevaux and K. Sörensen, “A genetic algorithm for robust schedules in a one-machine environment with ready times and due dates,” 4OR, vol. 2, no. 2, pp. 129–147, Jul. 2004.

[20] A. Yin and J. Wang, “Mixed Dispatch Rule for Single Machine Total Weighted Tardiness Problem,” J. Appl. Sci., vol. 13, no. 21, pp. 4616–4619, 2013.

[21] J. J. Kanet and X. Li, “A weighted modified due date rule for sequencing to minimize weighted tardiness,” J. Sched., vol. 7, no. 4, pp. 261–276, 2004.

[22] S. H. Yoon and I. S. Lee, “New constructive heuristics for the total weighted tardiness problem,” J. Oper. Res. Soc., vol. 62, no. 1, pp. 232–237, 2011.

ICU PATIENT DETERIORATION PREDICTION: A DATA-MINING APPROACH

Noura AlNuaimi, Mohammad M Masud and Farhan Mohammed

College of Information Technology,
United Arab Emirates University, Al-Ain, UAE
{noura.alnuaimi, m.masud, 200835338}@uaeu.ac.ae

ABSTRACT

A huge amount of medical data is generated every day, which presents a challenge in analysing these data. The obvious solution to this challenge is to reduce the amount of data without information loss. Dimension reduction is considered the most popular approach for reducing data size and also to reduce noise and redundancies in data. In this paper, we investigate the effect of feature selection in improving the prediction of patient deterioration in ICUs. We consider lab tests as features. Thus, choosing a subset of features would mean choosing the most important lab tests to perform. If the number of tests can be reduced by identifying the most important tests, then we could also identify the redundant tests. By omitting the redundant tests, observation time could be reduced and early treatment could be provided to avoid the risk. Additionally, unnecessary monetary cost would be avoided. Our approach uses state-of-the-art feature selection for predicting ICU patient deterioration using the medical lab results. We apply our technique on the publicly available MIMIC-II database and show the effectiveness of the feature selection. We also provide a detailed analysis of the best features identified by our approach.

KEYWORDS

Big data analytics; data mining; ICU; lab test; feature selection; learning algorithm

1. INTRODUCTION

Healthcare is changing from traditional medical practice to modern evidence-based healthcare. Evidence is based on patient data, which are collected from different resources like electronic health record (EHR) systems, monitoring devices and sensors [1]. One specific example of these technological advances is the observation and monitoring technologies for intensive care unit (ICU) patients. Currently, the data generated in the process of medical care ICUs are huge, complex and unstructured. Such data can be called big data due to their complexity, large size and difficulty to process in real-time [2]. However, these data could be used with the help of intelligent systems, such as big data analytics and decision support systems, to determine which patients are at an increased risk of death. This could support making the right decision to enhance the efficiency, accuracy and timeliness of clinical decision making in the ICU.

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 209–225, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51517

Reducing the amount of data without losing information is a great challenge. Dimension reduction would be the first solution to eliminate duplicate, useless and irrelevant features. In this paper, our goal is to propose an efficient mining technique to reduce the observation time in ICUs by predicting patient deterioration in its early stages through big data analytics. Our proposed technique has several contributions. First, we use the lab test results to predict patient deterioration. To the best of our knowledge, this is the first work that primarily uses medical lab tests to predict patient deterioration. Lab test results have a crucial role in medical decision making. Second, we identify most important medical lab tests using state-of-the-art feature-selection techniques without using any informed domain knowledge. Finally, our approach helps reduce redundant medical lab tests. Thus, healthcare professionals could focus on the most important lab tests to assist them, which would save not only costs but also valuable time in recovering the patient from a critical condition.

The paper is organised as follows. Section 2 presents the related work of predicting ICU death, Section 3 gives background on data mining and big data analytics, Section 4 illustrates our proposed approach, Section 5 summarises the MIMIC II dataset, Section 6 illustrates the experiment's work, Section 7 discusses the findings, and finally, the conclusion of this research is presented in Section 8.

2. LITERATURE REVIEW

This section reviews related works for predicting ICU death or the deterioration of ICU patients. We highlight some similarities and differences between some of the related works and the proposed work.

In [3], the authors developed an integrated data-mining approach to give early deterioration warnings for patients under real-time monitoring in the ICU and real-time data sensing (RDS). They synthesised a large feature set that included first- and second-order time-series features, detrended fluctuation analysis (DFA), spectral analysis, approximative entropy and cross-signal features. Then, they systematically applied and evaluated a series of established data-mining methods, including forward feature selection, linear and nonlinear classification algorithms, and exploratory under sampling for class imbalance. In our work, we are using the same dataset. However, we are using only the medical lab tests. Also, in our approach, we depend on feature selection to reduce the size of the dataset.

A health-data search engine was developed in [4] that supported predictions based on the summarised clusters patient types which claimed that it was better than predictions based on the non-summarised original data. In our work, we use only the medical lab tests, and we attempt to highlight the most important medical labs.

Liu et al. [4] investigated the critical feature size dimension. In their work, an ad hoc heuristic method based on feature-ranking algorithms was used to perform the experiment on six datasets. They found that the heuristic method is useful in finding the critical feature dimension for large datasets. In our work, we also use the ranking to rank the most useful features. However, we attempt to investigate the percentage of selected features that would be enough to have moderate model accuracy.

A survey of feature selection is presented in [6]. The authors presented a basic taxonomy of feature-selection techniques and discussed their use, variety and potential in a number of common and upcoming bioinformatics applications.

Cismondi et al. [5] proposed reducing unnecessary lab testing in the ICU. They applied artificial intelligence to study the predictability of future lab test results for gastrointestinal bleeding. This work is the closest work to our research; they have the same objective of reducing unnecessary lab tests. However, they only focus on gastrointestinal bleeding. In our work, we are targeting all cases in the ICUs.

3. BACKGROUND ON DATA MINING AND BIG DATA ANALYTICS

Healthcare, like other sectors, is facing the need for analysing large amounts of information, otherwise known as big data, which has become a major driver of innovation and success. Big data has potential to support a wide range of medical and healthcare functions, including clinical decision support [2].

Data mining is the analysis step of knowledge discovery. It is about the 'extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amount of data [10]'. When mining massive datasets, two of the most common, important and immediate problems are sampling and feature selection. Appropriate sampling and feature selection contribute to reducing the size of the dataset while obtaining satisfactory results in model building [4].

3.1. Feature Selection

In machine learning, feature selection or attribute selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used (a) to avoid overfitting and improve model performance, i.e. predict performance in the case of supervised classification and better cluster detection in the case of clustering, (b) to provide faster and more cost-effective models and (c) to gain deeper insight into the underlying processes that generated the data. In the context of classification, feature selection techniques can be organized into three categories, depending on how they perform the feature selection search to build the classification model: filter methods, wrapper methods and embedded methods, presented in table 1 [6] [7]:

1) Filter Methods are based on applying a statistical measure to assign a scoring to each feature. Then, features are ranked by score and either selected or removed from the dataset. The methods are often univariate and consider the feature independently or with regard to the dependent variable.
2) Wrapper Methods are based on the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy.
3) Embedded Methods are based on learning which features most contribute to the accuracy of the model while the model is being created.

Table 1: Feature selection categories.

Model Search	**Advantages**	**Disadvantages**
Filter	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier
Wrapper	Simple Interacts with the classifier Models feature decencies Less computational	Risk for overfitting More prone than randomized algorithms Classifier-dependent selection
Embedded	Interacts with the classifier More computational Models feature dependencies	Classifier-dependent selection

3.2. Data Classification Techniques

Classification is a pattern-recognition task that has applications in a broad range of fields. It requires the construction of a model that approximates the relationship between input features and output categories [8]. Some of the most popular techniques are discussed here in brief, all of which are used in our work.

1) The Naïve Bayes classifier is based on applying Bayes' theorem with strong independence assumptions between the features. As one of its main features, the Naïve Bayes classifier is easy to implement because it requires a small amount of training data in order to estimate the parameters, and good results can be found in most cases. However, it has class conditional independence, meaning it causes losses of accuracy and dependency [9].
2) Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines [10]. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets [11].
3) The ZeroR classifier simply predicts the majority category, which relies on the target and ignores all predictors. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [10].
4) A decision tree (J48) is a fast algorithm to train and generally gives good results. Its output is human readable, therefore one can see if it makes sense. It has tree visualizers to aid understanding. It is among the most used data mining algorithms. The decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points [10].
5) A RandomForest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [12].

4. PROPOSED APPROACH

In this section we introduce our approach for the Big Data mining technique for predicting ICU patient deterioration. Figure 1 shows the architecture of the proposed technique.

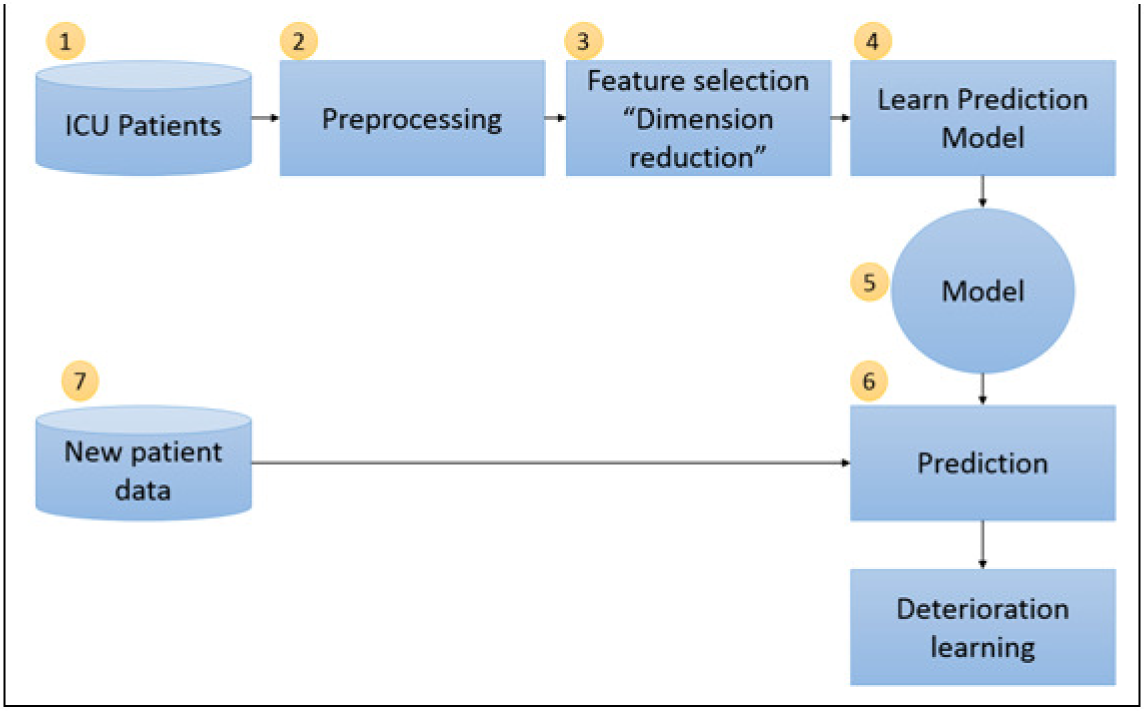


Figure 1: Architecture of the proposed approach

The data are collected from the database of ICU patients (step 1). Then the data are integrated, cleaned and relevant features are extracted (step 2). After that, feature selection or dimensionality reduction techniques are applied to obtain the best set of features and reduce the data dimension (step 3). Then the prediction model is learned using a machine learning approach (step 4). When a new patient is admitted to the CPU, the patient's data are collected incrementally (step 5). The patient data are evaluated by the prediction model (step 6) to predict the possibility of deterioration of the patient, and warnings are generated accordingly. Each of these steps is summarized here, and more details of the dataset are given in Section 5.

1) ICU Patient Data: The details of the data and the collection process are discussed in Section 5.
2) Preprocessing: At the preprocessing stage, we used two different datasets. These datasets were generated from a Labevents table. The first dataset contained the average value of applied medical tests, and the second contained the total number of times for each test was applied.
3) Feature Selection / Dimension Reduction: attribute selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. The goal here is to reduce the attributes so medical professional can identify the most important medical lab tests used by reducing the redundant tests. In our work, we select filter methods because they are moderately robust against the overfitting problem, as follows:
    a. Attribute evaluator: InfoGrainAttributeEval
    b. Search method: Ranker
    c. Attribute selection mode: use full training set
4) Learning: In our experiment we use a classification technique and five of the most popular classifier techniques: Naïve Bayes classifier, Support vector machine (SVM),

ZeroR classifier, decision tree (J48) and RandomForest. We use different types of machine learning order to avoid random results.

5) Model: The developed model aims to predict ICU patient deterioration by mining lab test results. Thus, observation time can be reduced in the ICUs and more actions can be taken in the early stages.
6) New patient data: When a new patient is admitted to the ICU, all his information is stored in the database. Some of these are incremental, such as vital sign readings, lab test results, medication events etc. The data of the patient again go through the preprocessing and feature extraction phases before they can be applied to the model.
7) Prediction: After each new test result, medication event, etc., the patient data are preprocessed and features are extracted to supply to the prediction model. The model predicts the probability of deterioration for the patient. This probability may change when new data (e.g. more test results) are accumulated and applied to the model. When the deterioration probability reaches a certain threshold specified by the healthcare providers, a warning is generated. This would help the healthcare providers to take proactive measures to save the patient from getting into a critical or fatal condition.

5. MIMIC II DATABASE

The MIMIC-II database is part of the Multiparameter Intelligent Monitoring in Intensive Care project funded by the National Institute of Biomedical Imaging and Bioengineering at the Laboratory of Computational Physiology at MIT, which was collected from 2001 to 2008 and represents 26,870 adult hospital admissions. In our work, we use MIMIC-II version 2.6 because is more stable than the newer version 3, which is still in the beta phase and needs further work of cleaning, optimizing and testing. MIMIC-II consists of two major components: clinical data and physiological waveforms.

The MIMIC dataset has three main features: (1) it is public; (2) it has a diverse and very large population of ICU patients; and (3) it contains high temporal resolution data, including lab results, electronic documentation, and bedside monitor trends and waveforms[13]. Several works have used the MIMIC dataset, such as [14], [15] and [16].

In our work, we focus on the clinical data, the LABEVENTS and LABITEMS tables. The Labevents table contains data of each patient's ICU stay, as presented in table 2, and table 3 contains descriptions of the lab events. Considering medical lab choice was done because we wanted to investigate the relationship between medical lab tests and patient deterioration so we could identify which medical tests have a major effect on clinical decision making. For example, the following information is about a patient who was staying at the ICU and was given a medical test. The following information was recorded at that time:

- Subject_ID: 2
- Hadm_ID: 25967
- IcuStay_ID: 3
- ItemID: 50468
- Charttime: 6/15/2806 21:48
- Value: 0.1
- ValueNum: 0.1
- Flag: abnormal

- ValueUOM: K/uL

Table 2: Labevents Table Description

Name	Type	Null	Comment
SUBJECT_ID	NUMBER(7)	N	Foreign key, referring to a unique patient identifier
HADM_ID	NUMBER(7)	Y	Foreign key, referring to the hospital admission ID of the patient
ICUSTAY_ID	NUMBER(7)	Y	ICU stay ID
ITEMID	NUMBER(7)	N	Foreign key, referring to an identifier for the laboratory test name
CHARTTIME	TIMESTAMP(6) WITH TIME ZONE	N	The date and time of the test
VALUE	VARCHAR2(100)	Y	The result value of the laboratory test
VALUENUM	NUMBER(38)	Y	The numeric representation of the laboratory test if the result was numeric
FLAG	VARCHAR2(10)	Y	Flag or annotation on the lab result to compare the lab result with the previous or next result
VALUEUOM	VARCHAR2(10)	Y	The units of measurement for the lab result value

Table 3: Labitems Table

Name	Type	Null	Comment
ITEMID	NUMBER(7)	N	Table record unique identifier, the lab item ID
TEST_NAME	VARCHAR2(50)	N	The name of the lab test performed
FLUID	VARCHAR2(50)	N	The fluid on which the test was performed
CATEGORY	VARCHAR2(50)	N	Item category
LOINC_CODE	VARCHAR2(7)	Y	LOINC code for lab item
LOINC_DESCRIPTION	VARCHAR2(100)	Y	LOINC description for lab item

6. EXPERIMENTS

We conducted four experiments to fulfil the different approaches to reach our goal of predicting ICU patient deterioration by mining lab test results. In each experiment, a different dataset resulted from pre-processing the MIMIC II v2.6 database.

6.1. Experiment 1: Building a Baseline of the Medical Lab Tests Average

1) Experiment Goal: The goal of this experiment was to investigate the effect of lab testing on predicting patient deterioration. Usually, medical professionals compare the result of the lab test with a reference range [17]. If the value is not within this range, the patient may face fatal consequences. Thus, the patient is kept under observation and the test is repeated again during a specific period. In our experiment, we investigated the average value of the same repeated test and, more precisely, how the average value of lab results could assist medical professionals in evaluating patient status.

Since we dealt with real cases, the only way to assess the quality and characteristics of a data mining model was through the final status of the patient, i.e. whether the patient survived or not. Thus, our evaluation criterion was how accurately our approach could predict whether the patient died or not.

2) Building the Dataset: The dataset was constructed by taking the average test result of each patient for each kind of test and make it one attribute. Thus one patient would be represented as one instance having 700 attributes, one for each test. If a test was not done, then the value of that attribute would be 0.
For example, the first patient record in the dataset would look like this:

P_ID	Avg1	Avg2	.....	Avg700	Dead/Alive
1	5.3	10		0	D

3) Pre-processing: After building the dataset, some values could not be reported because they were in text format. We used default values for these types of data. The total number of attributes was 619 with 2900 instances.

4) Base learners: In our experiment we used five classification algorithms to construct the model, namely NaiveBayes, SMO, ZeroR, J48 and RandomForest.

5) Evaluation: For a performance measurement, we did a 10-fold cross-validation of the dataset, and the confusion matrix was obtained to estimate four measures: accuracy, sensitivity, specificity and F-measure. As a result, RandomForest had the highest accuracy of 77.58%, followed by SMO with 76.86%, J48 with 75.27%, ZeroR with 70.24% and NavieBayes with 42.96%, as shown in Table 4. RandomForest and SMO have the same F-measures. The reason for the best performance by RandomForest is that it works relatively well when used with high-dimensional data with a redundant/noisy set of features [12]

Table 4: Experiment 1 results

Algorithm	**Learning Machine**	**Detailed Accuracy**			
		Accuracy	**Precision**	**Recall**	**F-Measure**
Bayes	NavieBayes	42.96%	0.672	0.430	0.404
Functions	SMO	76.86 %	0.759	0.769	0.762
Rule	ZeroR	70.24 %	0.493	0.702	0.580
Tree	J48	75.27%	0.749	0.753	0.751
Tree	RandomForest	77.58 %	0.765	0.776	0.762

6.2. Experiment 2: Average Medical Lab Tests Feature Selection

1) Experiment Goal: The goal of this experiment was to study the relationship between feature selection and classification accuracy. Feature selection is one of the dimensionality reduction techniques for reducing the attribute space of a feature set. More precisely, it determines how many features should be enough to give moderate accuracy.

2) Building the Dataset: In this experiment we used the same dataset that we used in experiment 1.

3) Pre-processing: In this experiment we built ten datasets depending on the number of selected features. We start with the first dataset, which contained only 10% of the total attributes. Then each time, we increased the total feature selections by 10%. For example, dataset 1 contains 10% of the total attributes, dataset 2 contains 20% of the total attributes, dataset 3 contains 30% of the total attributes and so on till dataset 10 contains all 100% of the total attributes.

   For feature selection, we use supervised.attribute. InfoGainAttributeEval from WEKA. This filter is a wrapper for the Weka class that computes the information gain on a class [18].

   - Attribute Subset Evaluator: InfoGainAttributeEval
   - Search Method: Ranker.
   - Evaluation mode: evaluate all training data

4) Base learner: After generating all of the reduced datasets, we use the J48 algorithm to construct a model.

5) Evaluation: For each reduced dataset, we applied 10-fold cross-validation for evaluating the accuracy. Table V shows the results in numbers, and Figure 2 shows them as a chart. The results indicate that taking only the most related 10% of the total features can give a 75.10% accurate result, which is comparable to the accuracy of the full feature set. This indicates that not all of the features are required to get the highest accuracy. However, there are some fluctuations, such as at 20%, the accuracy drops a little. We conclude that selecting 50 to 80% of the attributes should give moderately satisfying accuracy.

Table 5: Experiment 2 Feature selection.

% of Features Selected	**# of Features Selected**	**Detailed Accuracy**		
		Accuracy	**Number of leaves**	**Size of the Tree**
10%	62	75.10%	200	399
20%	124	73.59%	201	401
30%	186	75.10%	185	369
40%	248	74.93%	179	357
50%	310	75.17%	189	377
60%	371	74.79%	187	373
70%	433	75.00%	189	377
80%	495	75.31%	184	367
90%	557	74.97%	183	365
100%	619	74.86%	184	367

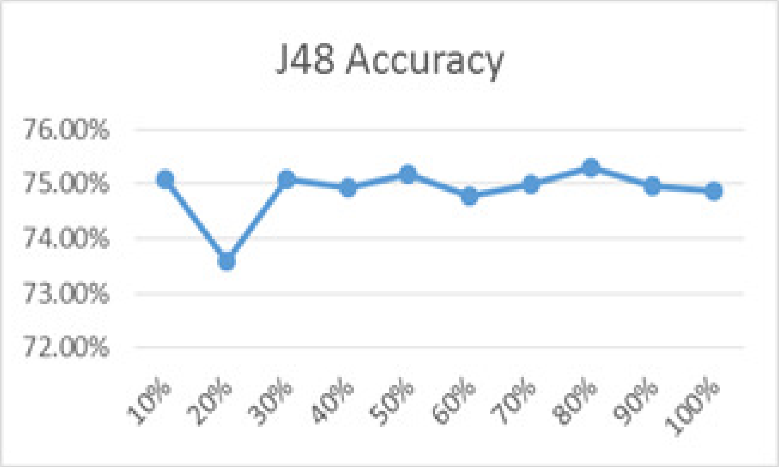


Figure 2: Average datasets accuracy.

6.3. Experiment 3: Building a Baseline for the Total Number of Medical Lab Tests

1) Experiment Goal: The goal of this experiment was to investigate the effect of the total number of lab tests conducted on predicting patient deterioration. Usually, medical professionals keep requesting the same medical test over a brief period to compare the result with a reference range [17]. If the value is not within the range, it means the patient may be in danger, so the test is repeated again and again. Our goal was to predict at what total number a medical professional should start immediate action and, more precisely, how the total number of medical lab tests could assist the medical professional in evaluating the patient's status.

2) Building the Dataset: The dataset was built by taking the total number of tests taken for each patient for each type of test and make it one attribute. Then one patient would be represented as one instance having 700 attributes, one for each test. If a test was not done, then the value of that attribute would be 0.

   For example, the dataset would look like this:

P_ID	Count1	Count2	…	Count700	Dead/Alive
1	5	0		1	D

3) Pre-processing: The dataset was randomized first, then two datasets were generated, Count_Training_Validation_Dataset and Count_testing_Dataset. This step was repeated ten times because we used randomization to distribute the instances between the two datasets.

4) Base learners: Five learning algorithms were used to build the model, namely NaiveBayes, SMO, ZeroR, J48 and RandomForest.

5) Evaluation: The training data were first used to build the model and then evaluated using a percentage split via test data. For a performance measurement, the confusion matrix was obtained to estimate four measures: accuracy, sensitivity, specificity and F-measure. Table 6 shows that SMO and RandomForest have almost equal levels of accuracy, around 75%. Even after testing the model with the test datasets, SMO and RandomForest still have the highest

accuracy among the other techniques. The reason for this higher accuracy is that the amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets [11].

Table 6: Experiment 3 results.

Algorithm	**Learning Machine**	**Detailed Accuracy**			
		Accuracy	**Precision**	**Recall**	**F-Measure**
Bayes	NavieBayes	73.66%	0.718	0.737	0.713
Funtions	SMO	75.44%	0.739	0.755	0.723
Rule	ZeroR	70.46%	0.497	0.705	0.583
Tree	J48	73.16%	0.728	0.732	0.692
Tree	RandomForest	75.73%	0.742	0.757	0.739

Table 7: Experiment 3 Results

Algorithm	**Learning Machine**	**Detailed Accuracy**			
		Accuracy	**Precision**	**Recall**	**F-Measure**
Bayes	NavieBayes	73.48%	0.716	0.735	0.711
Funtions	SMO	74.85%	0.737	0.749	0.716
Rule	ZeroR	69.72%	0.486	0.697	0.573
Tree	J48	72.44%	0.722	0.724	0.723
Tree	RandomForest	75.30%	0.739	0.753	0.736

6.4. Experiment 4: Feature Selection for Total Number of Medical Lab Tests

1) Experiment Goal: The goal of this experiment was to study the relationship between feature selection and classification accuracy. Feature selection is one of the dimensionality reduction techniques for reducing the attribute space of a feature set. More precisely, it measures how many features should be enough to give moderate accuracy.

2) Building the Dataset: In this experiment we used a count dataset.

3) Pre-processing: In the pre-processing step, we built ten datasets depending on the number of selected features. The first dataset contained only 10% of the total attributes. Then we increased the total feature selections by 10% with each new dataset. For example, dataset 1 contained 10% of the total attributes, dataset 2 contained 20% of the total attributes, dataset 3 contained 30% of the total attributes and so on till dataset 10 contained all 100% of the total attributes.

4) For feature selection, we used supervised.attribute. InfoGainAttributeEval from WEKA. This filter is a wrapper for the Weka class that computes the information gain on a class [18].

- Attribute Subset Evaluator: InfoGainAttributeEval
- Search Method: Ranker.
- Evaluation mode: evaluate on all training data

5) Base learner: After generating all reduced datasets, we used the J48 algorithm as a base learner.

6) Evaluation: Each feature-reduced dataset went through a 10-fold cross-validation for evaluation. Figure 3 shows the accuracy of all count datasets. The detail values are also reported in Table 4. From the results we observe that selecting 60 to 70% of the attributes gives the highest accuracy. This also concludes that all features (i.e., lab tests) may not be necessary to attain a highly accurate prediction of patient deterioration.

Table 8: Experiment 4 Results

% of Features Selection	**# of Features Selection**	**Detailed Accuracy**		
		Accuracy	**Number of leaves**	**Size of the Tree**
10%	62	71.45%	237	473
20%	124	73.90%	250	499
30%	186	73.55%	247	493
40%	248	72.79%	252	503
50%	310	73.41%	252	503
60%	371	73.66%	254	507
70%	433	74.24%	254	507
80%	495	74.10%	254	507
90%	557	74.14%	265	529
100%	619	73.59%	259	517

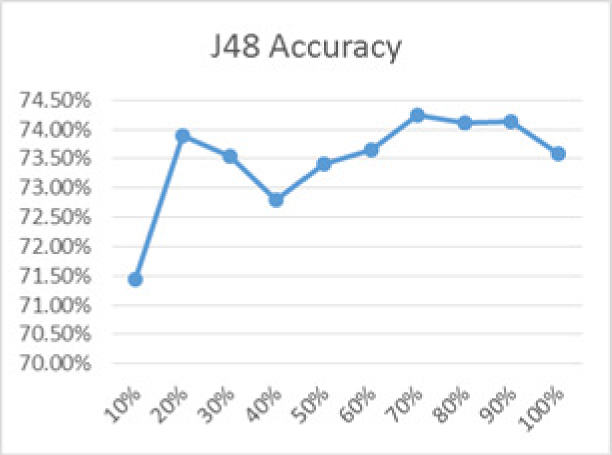


Figure 3: Count dataset accuracy.

7. DISCUSSION

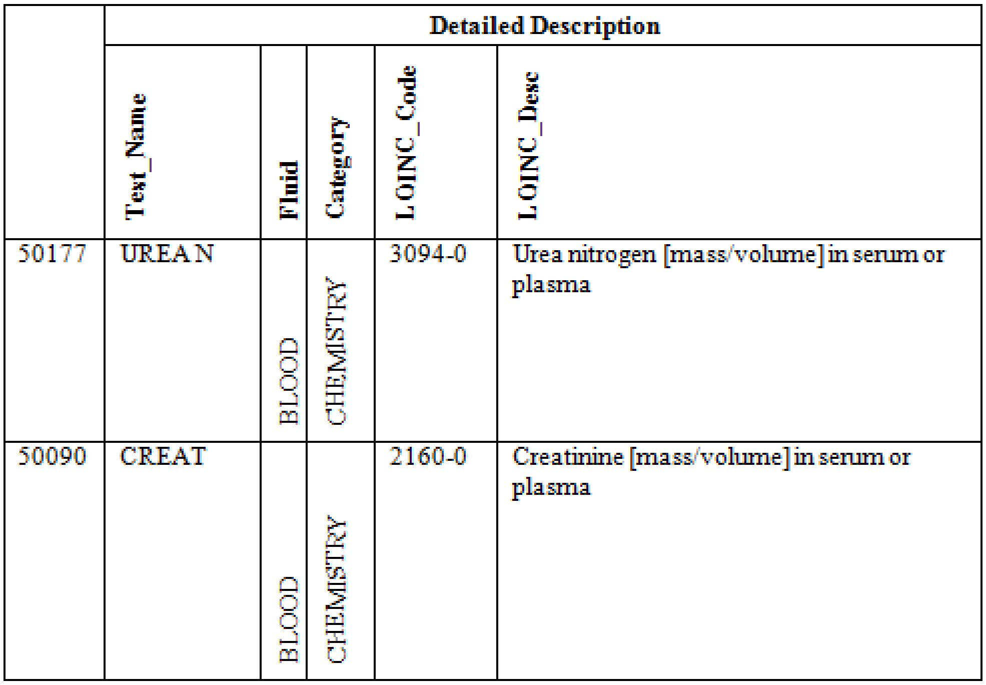
It should be noted that the feature selections were done without any domain knowledge and without any intervention from medical experts. However, in the analysis we would like to emphasize the merit of feature selection in choosing the best tests, which could be further verified and confirmed by a medical expert.

First we compare the selected features selected from the two datasets, namely the average dataset and the count dataset. Table 9 shows the 10 best features chosen by the two approaches and highlights the common lab tests between the two approaches (i.e. using the average of tests and count of tests). Table 10 shows more details about the common tests.

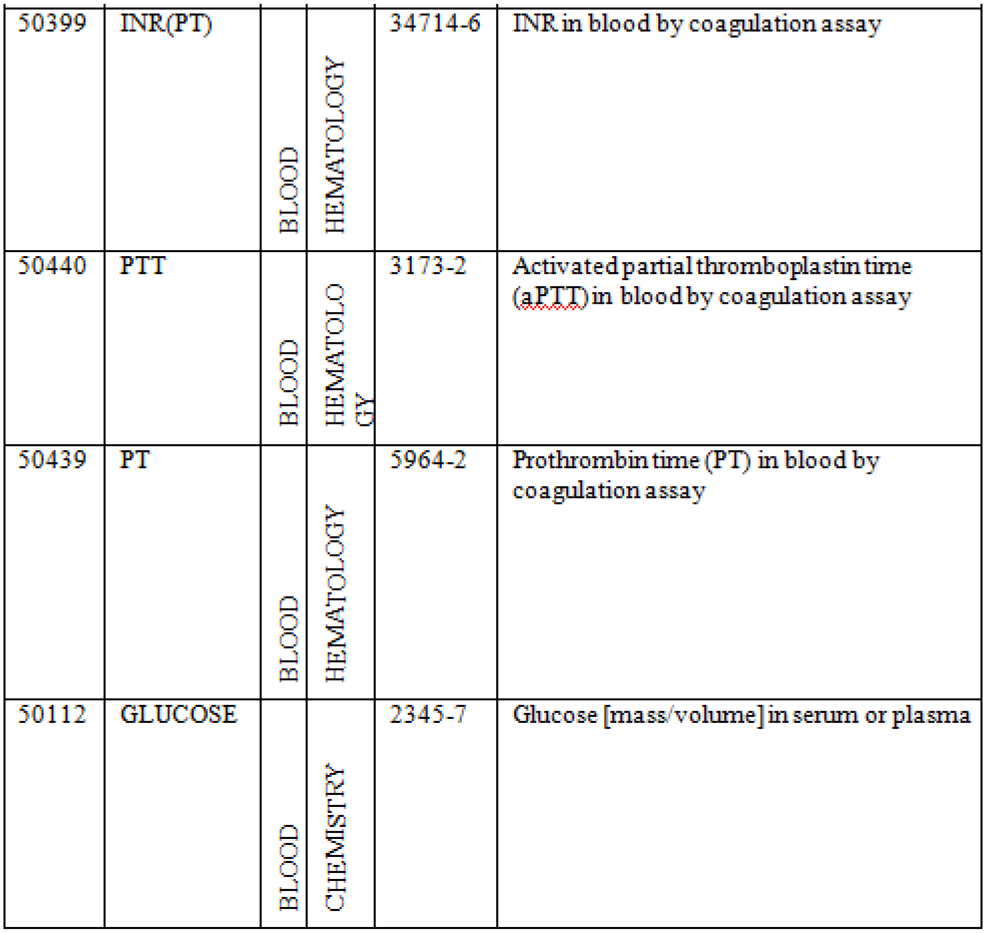
Table 9: Final Results

	Detailed Accuracy	
	Average Dataset	Count Dataset
Best ranked 10 from the 10% of selected features	50177 50090 50060 50399 50386 50440 50408 50439 50112 50383	50148 50112 50140 50399 50177 50439 50090 50440 50079 50068

Table 10: Medical Lab Test Details.



	Detailed Description				
	Test_Name	Fluid	Category	LOINC_Code	LOINC_Desc
50177	UREA N	BLOOD	CHEMISTRY	3094-0	Urea nitrogen [mass/volume] in serum or plasma
50090	CREAT	BLOOD	CHEMISTRY	2160-0	Creatinine [mass/volume] in serum or plasma

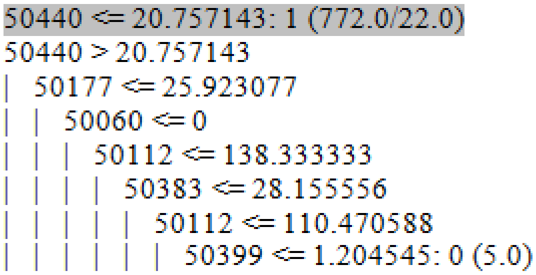


50399	INR(PT)	BLOOD	HEMATOLOGY	34714-6	INR in blood by coagulation assay
50440	PTT	BLOOD	HEMATOLOGY	3173-2	Activated partial thromboplastin time (aPTT) in blood by coagulation assay
50439	PT	BLOOD	HEMATOLOGY	5964-2	Prothrombin time (PT) in blood by coagulation assay
50112	GLUCOSE	BLOOD	CHEMISTRY	2345-7	Glucose [mass/volume] in serum or plasma

LOINC is an abbreviation for logical observation identifiers names and codes. LOINC is clinical terminology important for laboratory test orders and results [19]. ARUP Laboratories [20] is a national clinical and anatomic pathology reference laboratory and a worldwide leader in innovative laboratory research and development. We used their web page and others to clarify more about the medical lab tests in table 10 as follows:

- UREAN (50177): This test is conducted using the patient's blood. This test is recommended to screen for kidney dysfunction in patients with known risk factors (e.g. hypertension, diabetes, obesity, family history of kidney disease). The panel includes albumin, calcium, carbon dioxide, creatinine, chloride, glucose, phosphorous, potassium, sodium and BUN and a calculated anion gap value. Usually, the result is reported within 24 hours [20].
- CREAT (50090): This test is conducted using the patient's blood. It is a screening test to evaluate kidney function [20].
- INR(PT) (50399): This test is conducted using the patient's blood by coagulation assay [13].
- PTT (50440): This test is carried out to answer two main questions: does the patient have antiphospholipid syndrome (APLS), and does the patient have von Willebrand disease? If so, which type? It is carried out by mechanical clot detection [21].
- PT (50439): This test is conducted using the patient's blood by coagulation assay [13].

- GLUCOSE (50112): This test is used to check glucose, which is a common medical analytic measured in blood samples. Eating or fasting prior to taking a blood sample has an effect on the result. Higher than usual glucose levels may be a sign of prediabetes or diabetes mellitus [22].
- The result of the top 10 selected features from the average dataset allows us to build a model using decision tree J48. This model would allow a medical professional to predict the status of a patient in the ICU as follows:



```
50440 <= 20.757143: 1 (772.0/22.0)
50440 > 20.757143
|   50177 <= 25.923077
|   |   50060 <= 0
|   |   |   50112 <= 138.333333
|   |   |   |   50383 <= 28.155556
|   |   |   |   |   50112 <= 110.470588
|   |   |   |   |   |   50399 <= 1.204545: 0 (5.0)
```

For example, if the lab test (name: PTT, ID 50440, LOINC: 3173-2) result value is <= 20.757143, then the probability is very high (772.0/22.0~ 97.2%) that the patient is going to die (class:1). This model has 78.6897% overall accuracy.

8. Conclusion and Future Work

In this paper, we presented our proposed approach to reduce the observation time in the ICU by predicting patient deterioration in its early stages. In our work, we presented experiments 1 and 3 to build a model to predict patient deterioration. Experiments 2 and 4 identified the most important medical lab tests, then highlighted the common tests between the two datasets. The four experiments would help medical professionals to take better decisions in a very short time.

For future work, the authors are planning to carry out more experiments using bigger data. Big data analytics would bring potential benefits to support taking the right decision to enhance the efficiency, accuracy and timeliness of clinical decision making in the ICU.

References

[1] "Big Data in Healthcare: Intensive Care Units as a Case Study." [Online]. Available: http://ercim-news.ercim.eu/en97/ri/big-data-in-healthcare-intensive-care-units-as-a-case-study. [Accessed: 28-Aug-2015].

[2] Raghupathi, Wullianallur and Raghupathi, Viju, "Big data analytics in healthcare: promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, p. 3, 2014.

[3] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," in Knowledge discovery and data mining, 2012, pp. 1140–1148.

[4] Q. Liu, Sung, Andrew H, Ribeiro, Bernardete, and Suryakumar, Divya, "Mining the Big Data: The Critical Feature Dimension Problem," Adv. Appl. Inform. IIAIAAI 2014 IIAI 3rd Int. Conf. On, pp. 499–504, 2014.

[5] Federico Cismondi, Leo A. Celi, André S. Fialho, Susana M. Vieira, Shane R. Reti, Joao MC Sousa, and Stan N. Finkelstein, "Reducing unnecessary lab testing in the ICU with artificial intelligence," Int. J. Med. Inf., vol. 82, no. 5, pp. 345–358, 2013.

[6] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga, "A review of feature selection techniques in bioinformatics," bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[7] "An Introduction to Feature Selection - Machine Learning Mastery." [Online]. Available: http://machinelearningmastery.com/an-introduction-to-feature-selection/. [Accessed: 06-Sep-2015].

[8] S. Bouktif et al, "Ant Colony Optimization Algorithm for Interpretable Bayesian Classifiers Combination: Application to Medical Predictions," PLoS ONE, vol. 9, no. 2, 2014.

[9] X. Wu et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

[10] Chitra Nasa and Suman, "Evaluation of Different Classification Techniques for WEB Data," Int. J. Comput. Appl., vol. 52, no. 9, 2012.

[11] John C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Adv. Kernel Methods—support Vector Learn., vol. 3, 1999.

[12] Leo Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[13] "MIMIC II Database." [Online]. Available: https://mimic.physionet.org/database.html. [Accessed: 20-Aug-2015].

[14] Lee J, Govindan S, Celi L, Khabbaz K, and Subramaniam B, "Customized prediction of short length of stay following elective cardiac surgery in elderly patients using a genetic algorithm," World J Cardiovasc Surg, vol. 3, no. 5, pp. 163–170, Sep. 2013.

[15] Lehman LH, Saeed M, Talmor D, Mark R, and Malhotra A, "Methods of blood pressure measurement in the ICU," Crit Care Med, vol. 41, no. 1, pp. 34–40, 2013.

[16] Lehman L, Long W, Saeed M, and Mark R, "Latent topic discovery of clinical concepts from hospital discharge summaries of a heterogeneous patient cohort," in Proceedings of the 36th International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.

[17] "Laboratory Test Reference Ranges | Calgary Laboratory Services." [Online]. Available: https://www.calgarylabservices.com/lab-services-guide/lab-reference-ranges/. [Accessed: 03-Sep-2015].

[18] "Feature Selection Package Documentation." [Online]. Available: http://featureselection.asu.edu/documentation/infogain.htm. [Accessed: 04-Sep-2015].

[19] "LOINC Codes - Mayo Medical Laboratories." [Online]. Available: http://www.mayomedicallaboratories.com/test-catalog/appendix/loinc-codes.html. [Accessed: 10-Sep-2015].

[20] "ARUP Laboratories: A National Reference Laboratory." [Online]. Available: http://www.aruplab.com/. [Accessed: 10-Sep-2015].

[21] "UCSF Departments of Pathology and Laboratory Medicine | Lab Manual | Laboratory Test Database | Activated Partial Thromboplastin Time." [Online]. Available: http://labmed.ucsf.edu/labmanual/db/data/tests/802.html. [Accessed: 10-Sep-2015].

[22] "2345-7." [Online]. Available: http://s.details.loinc.org/LOINC/2345-7.html?sections=Comprehensive. [Accessed: 10-Sep-2015].

AUTHORS

Noura Al Nuaimi is pursuing a PhD in Information Technology with Dr Mohammad Mehedy Masud at United Arab Emirates University (UAEU). She holds an MSc in Business Administration from Abu Dhabi University and a BSc in Software Engineering from UAEU. Her research interests focus on data mining and knowledge discovery, cloud computing, health information systems, search engines and natural language processing. She has published research papers in IEEE Computer Society and IEEE Xplore.

Dr Mohammad Mehedy Masud is currently an Assistant Professor at the United Arab Emirates University (UAEU). He joined the College of Information Technology at UAEU in spring 2012. He received his PhD from University of Texas at Dallas (UTD) in December 2009. His research interests are in data mining, especially data stream mining and big data mining. He has published more than 30 research papers in journals including IEEE Transactions on Knowledge and Data Engineering (TKDE), Journal of Knowledge and Information Systems (KAIS), ACM Transactions on Management Information Systems (ACM TMIS) and peer-reviewed conferences including IEEE International Conference on Data Mining (ICDM), European Conference on Machine Learning (ECML/PKDD) and Pacific Asia Conference on KDD. He is the principal inventor of a US patent application and lead author of the book "Data Mining Tools for

Malware Detection". Dr Masud has served as a program committee member of several prestigious conferences and has been serving as the official reviewer of several journals, including IEEE TKDE, IEEE TNNLS and DMKD. During his service at the UAEU he has secured several internal and external grants as PI and co-PI.

Farhan Mohammed is a graduate from the College of Information Technology in United Arab Emirates University specializing in Information Technology Management. He obtained his Bachelor's in Management Information Systems from United Arab Emirates University, Al Ain, UAE. He has worked under several professors and published four conference papers and a journal paper for IEEE sponsored conferences. Currently he is working as a research assistant in data mining in the health industry to develop models on health deterioration prediction. His area of interests lies in smart cities, UAVs, data mining, and image and pattern recognition.

INTENTIONAL BLANK

SENSITIVITY ANALYSIS IN A LIDAR-CAMERA CALIBRATION

Angel-Iván García-Moreno[1], José-Joel Gonzalez-Barbosa[1], Juan B. Hurtado-Ramos and Francisco-Javier Ornelas-Rodriguez.

[1]Research Center for Applied Science and Advanced Technology (CICATA), Instituto Politécnico Nacional (IPN). Cerro Blanco 141 Col. Colinas del Cimatario, Querétaro, México.
angelivan.garciam@gmail.com

ABSTRACT

In this paper, variability analysis was performed on the model calibration methodology between a multi-camera system and a LiDAR laser sensor (Light Detection and Ranging). Both sensors are used to digitize urban environments. A practical and complete methodology is presented to predict the error propagation inside the LiDAR-camera calibration. We perform a sensitivity analysis in a local and global way. The local approach analyses the output variance with respect to the input, only one parameter is varied at once. In the global sensitivity approach, all parameters are varied simultaneously and sensitivity indexes are calculated on the total variation range of the input parameters. We quantify the uncertainty behaviour in the intrinsic camera parameters and the relationship between the noisy data of both sensors and their calibration. We calculated the sensitivity indexes by two techniques, Sobol and FAST (Fourier amplitude sensitivity test). Statistics of the sensitivity analysis are displayed for each sensor, the sensitivity ratio in laser-camera calibration data

KEYWORDS

Sensitivity analysis, Remote sensing, LIDAR calibration, Camera calibration.

1. INTRODUCTION

Many papers address the calibration LiDAR-camera but very few papers address the uncertainty and sensitivity analysis in the data fusion between a laser sensor and a camera. The interest of this development is to present a study of the quantification of the variability in own methodology results according to the input uncertainties in the calibration model of a multi-sensor platform for urban dimensional reconstruction tasks.

Once the calibration model and data fusion are defined, the sensitivity analysis will determine the uncertainty in input parameters [13]. Many approaches to sensitivity analysis and uncertainty exist, including (1) based on sampling, that uses samples generation considering its probability distribution to analyze the results of its variation [7], (2) differential analysis, which is to approximate the model to a Taylor series and then performing an analysis of variance to obtain the sensitivity analysis [2], (3) Fourier amplitude sensitivity test (FAST), is based on the variation

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 227–239, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51518

of the predictions of the models and the contributions of individual variables to the variance [14], (4) response surface methodology (RSM) in which an experiment is designed to provide a reasonable response values of the model and then, determine the mathematical model that best fits. The ultimate goal of RSM is to establish the values of the factors that optimize the response (cost, time, efficiency, etc.) [11].

Generally, it can perform a sensitivity analysis in a local or global way [13]. The local approach analyses the output variance with respect to the input parameters. Input parameters are altered within a small range around a nominal value and a particular analysis of each parameter is performed. This analysis only provide information based where it is calculated, since no analyses the entire input parameter space. Furthermore, if the model is not continuous the analysis is unable to be performed. On the other hand, global sensitivity approach defines the uncertainty of the output to the uncertainty of the input factors, by sampling Probability Density Functions (PDF) associated with the input parameters. For this approach, all parameters are varied simultaneously and sensitivity indexes are calculated on the total variation range of the input parameters.

The Monte Carlo (MC) method provides a standard technique for assessing uncertainty in models and modeling data fusion; simulating and sampling the input variables [8,17]. The MC method generates pre-defined pseudo-random numbers with a probability distribution according to its *PDF*. The number of sequences to be simulated must be determined according to the recommendation in [4]. When a model contains too many variables, the uncertainty and sensitivity analysis using Monte Carlo method becomes difficult and with a high computationally cost. This difficulty arises because too many variables require a large number of simulations.

In the Latin Hypercube Sampling (LHS) method, parameters are treated as pseudo-strata and numbers are distributed in proportion to the elements of each strata sample. The *PDF* is generated from the average of each stratum and must have the same distribution of the elements in the sample strata to be calculated under predetermined conditions. The amount of pseudo-random numbers to be generated for the simulation is set similarly to that defined in MC. An analysis of the efficiency and speed of the LHS method against MC method is presented in [01].

Other methods for global sensitivity analysis are presented in [3, 9, 12].

2. Model Acquisition Platform

The goal is geometrically model our multi-sensors platform behaviour (Fig. 1). Consists by a *Velodyne HDL-64E* laser scanner, a *Point Grey Ladybug2* multi-camera system.

2.1. LiDAR model

To calculate the LiDAR extrinsic parameters, the methodology presented in [5] was implemented. The three-dimensional data is modeled as an *AllPointCloud* $= f(\boldsymbol{r}, \theta, \phi)$ with respect to origin, where $\boldsymbol{r}$ is the radius of the surface, θ colatitude or zenith angle, and ϕ azimuthal angle on the unit sphere. The development of the mathematical model for the captured 3D points is based on the transformation of the spherical coordinates for each position of the unitary sphere onto cartesian coordinates. The key point of this transformation is that every involved parameter includes a perturbation as defined in Eq. 1.

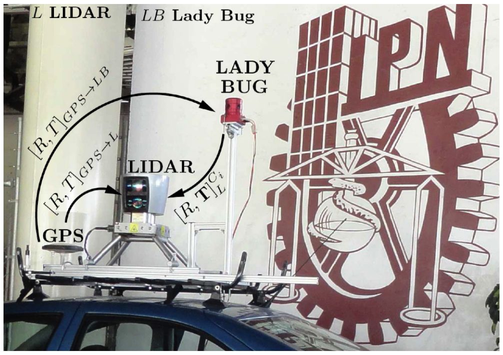


Figure 1 Sensor platform composed of LiDAR Velodyne HDL-64E (*L*), *Ladybug2* (*LB*) and GPS. *Ladybug* spherical digital video camera system has six cameras (*C*). $[\mathbf{R},\mathbf{T}]_{\mathbf{L}}^{\mathbf{C}}$ represent translation and rotation of the LiDAR and six camera frames of the *Ladybug2*.

$$\begin{aligned} x' &= d_x' \sin(\theta+\Delta\theta) - h_{osc} \cos(\theta+\Delta\theta)\,, \\ y' &= d_x' \cos(\theta+\Delta\theta) + h_{osc} \sin(\theta+\Delta\theta)\,, \\ z' &= (ds+\Delta ds)\sin(\phi+\Delta\phi) + v_{osc} \cos(\phi+\Delta\phi), \end{aligned} \tag{1}$$

where $\mathrm{d}_x' = (\mathrm{ds} + \Delta\, \mathrm{d}_s)\cos(\phi + \Delta\,\phi) - \mathrm{v}_{OSC} \sin\phi + \Delta\phi$, see Fig. 2.

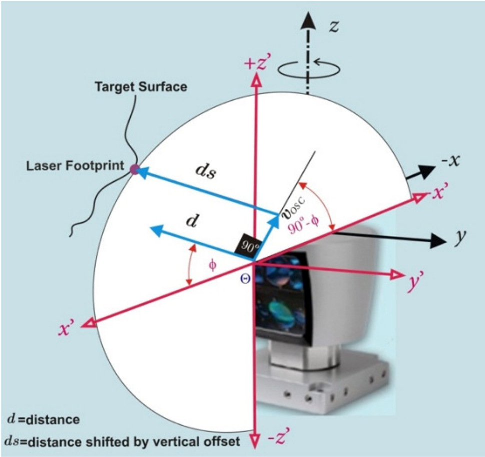


Figure 2 LiDAR Velodyne HDL-64E configuration.

2.2. Multi-camera model

To get the intrinsic parameters we followed the calibration methodology presented in [18]. A camera is modeled by the usual pinhole camera model: The image $\boldsymbol{u}^C$ of a 3D point $\boldsymbol{X}^L$ is formed by an optical ray from $\boldsymbol{X}^L$ passing through the optical center C and intersecting the image plane. The relationship between the 3D point $\boldsymbol{X}^L$ and its image projection $\boldsymbol{u}^C$ is given by Eq. 2.

$$s\hat{u}^C = A^C\left[R_L^C, T_L^C\right]\widehat{X}^L = P_L^C\widehat{X}^L \tag{2}$$

$$\text{with } A^C = \begin{bmatrix} -k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\text{and } A^C[R_L^C, \boldsymbol{T}_L^C] \, ,$$

where s is a scale factor, $[R_L^C, \boldsymbol{T}_L^C]$ are the extrinsic parameters and represent the rigid transformation between a point in the LiDAR frame L to the camera frame C, A^C is the intrinsic camera parameters matrix, where (u_o, v_0) are the principal point coordinates in the image, focal length *f*, $-k_u f = \alpha \; and \; k_v f = \beta$ are the image scale factors in the axis u y v. The P_L^C is the projection matrix.

2.3. LiDAR-camera calibration

The extrinsic transformation between the LiDAR and the camera frame was computed using Eq. 3.

$$[R_L^C, \boldsymbol{T}_L^C] = [R_L^C, \boldsymbol{T}_L^C]_W^C * \; ([R, \boldsymbol{T}]_W^L)^{-1} \tag{3}$$

where R and $\boldsymbol{T}$ represents de rotation and translation respectively, C and L are the camera and LiDAR and W represents the word frame.

The pattern acquired by the LiDAR is transformed onto the image frame using the extrinsic parameters $[R_L^{C_i}, \boldsymbol{T}_L^{C_i}]$. This transformation allows us to reference in the camera the points acquired by the LiDAR. The projection is completed using the intrinsic camera parameters A^C, Fig. 3.

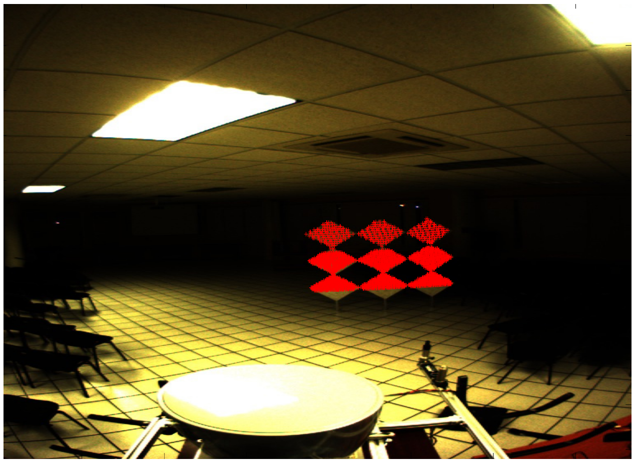


Figure 3 The red points are acquired by the LiDAR and projected onto the image.

3. Sensitivity Analysis

Two techniques were used for analysis, Sobol and FAST. The advantages that are: (1) parameters are evaluated throughout its range of variation, (2) estimating the expected values and the variance of each parameter is calculated directly, (3) calculate the contribution that each parameter has on the global sensitivity, (4) determine the effects of the interaction between the parameters in the sensitivity analysis, and (5) no modifications are required in the calibration model for the analysis.

3.1. Sobol

Suppose the model is given by $Y = f(X_1, X_2, \dots, X_k)$, where X_i are independent input parameters, Y is the model output. Using dispersion analysis by [1], f can be linear or nonlinear, and the sensitivity analysis evaluates the contribution of each parameter X_i in the variance of Y. This study is known as analysis of variance (ANOVA). For sensitivity indices for each parameter independently V variance model is decomposed as:

$$\mathrm{V} = \sum_{\mathrm{i}} \mathrm{V_i} + \sum_{\mathrm{i}<j} V_{i,j} + \sum_{\mathrm{i}<j<m} V_{i,j,m} + \cdots + \mathrm{V_k} \tag{4}$$

where $V_k = V\left(E\left(Y \middle| X_i, X_j, \dots, X_k\right)\right) - V_i - V_j - \cdots - V_k$. As a rule, the Eq. 4 has a total of terms $\sum\left(C_1^k + C_2^2 + \cdots + C_k^k\right) = 2^k - 1$. Sensitivity analysis indices are computed:

$$\mathrm{S}_{\mathrm{i}_1, i_2, \dots, i_k} = \frac{\mathrm{V}_{\mathrm{i}_1, i_2, \dots, i_k}}{\mathrm{V}} \tag{5}$$

where $i_1, i_2, \cdots, i_k$ are the input parameters. Then, all the sensitivity indexes that allow us to observe the interaction between inputs and nonlinearity must meet the condition:

$$\sum_{\mathrm{i}=1}^{\mathrm{k}} S_i + \sum_{i} \sum_{j>i} S_{i,j} + \cdots + S_k = 1 \tag{6}$$

On the other hand, the index $S_i = \frac{V_i}{V}$ only shows the effect of the parameter X_i in the model output, but not analyzes the interaction with other parameters. To estimate the total influence of each parameter, the total partial variance is calculated:

$$V_i^{tot} = \sum_{\mathrm{i}} V_{i_1, \dots, i_k} \tag{7}$$

the sum is over all the different groups of indexes that satisfy the condition $1 \le i_1 < i_2 < \cdots < i_k \le s$, where one of the indexes is equal to i. Then, the total sensitivity index is given by:

$$S_i^{tot} = \frac{\mathrm{V_i^{tot}}}{V} \tag{8}$$

The total sensitivity index (total variance) S_i^{tot} represents the expected percentage of the variance that remains in the model output if all the parameters are known except i. It follows then $0 \le S_i \le S_i^{tot} \le 1$. The result is S_i^{tot} and S_i indicates the interaction between parameter i and the other ones.

3.2. FAST

FAST allows generate independent sensitivity indexes for each input parameter using the same number of iterations. The method idea is convert Eq. 4 in a uni-dimensional integer s using the transformation functions G_i:

$$X_i = G_i(\sin \omega_i s) \tag{9}$$

where $s \in (-\pi, \pi)$. A good choice of the transformation G_i and frequency ω_i will assess the model in a sufficient number of points [14]. Then the expected value of Y can be approximated by:

$$E(Y) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)ds \tag{10}$$

by

$$f(s) = f(G_1(\sin \omega_1 s), \dots, G_k(\sin \omega_k s)) \tag{11}$$

So therefore, we can approximate the variance of Y as:

$$\begin{aligned} V(Y) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f^2(s)ds - [E(Y)]^2 \\ &\approx \sum_{i=-\infty}^{\infty} \left(A_i^2 + B_i^2\right) - (A_0^2 + B_0^2) \\ &\approx 2\sum_{i=1}^{\infty} \left(A_i^2 + B_i^2\right) \end{aligned} \tag{12}$$

where A_i and B_i are the Fourier coefficients. The contribution of X_i in the variance $V(Y)$ can be approximated by:

$$D_i \approx \sum_{p=1}^{M} (A_{p\omega i}^2 + B_{p\omega i}^2) \tag{13}$$

where ωi is related to the value G_i in Eq. 11 by $p = 1,2 \cdots M$. Will then M, the maximum harmonic. The sensitivity coefficients by the FAST method for each parameter is calculated:

$$S_i = \frac{\mathrm{V_i}}{V(Y)} \doteq \frac{\sum_{p=1}^{M}(A_{p\omega i}^2 + B_{p\omega i}^2)}{\sum_{i=1}^{M}\left(A_i^2 + B_i^2\right)} \tag{14}$$

4. RESULTS

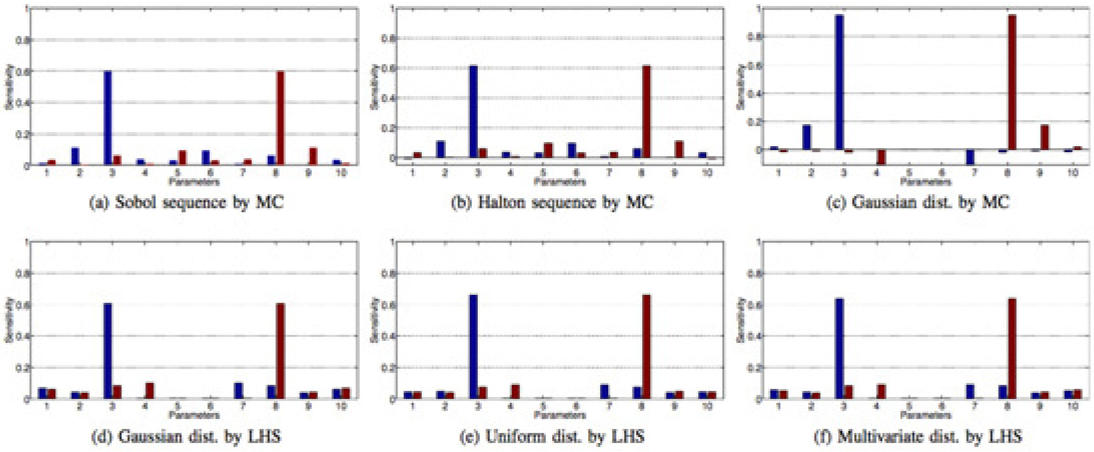
4.1. Sensitivity

4.1.1. Ladybug2

Were simulated *2,000* samples by MC and LHS. The camera was intrinsically calibrated by the method presented in [18]. Sobol was the first technique in which the camera sensitivity was assessed (section 3.1). The determination of global indexes (S_i) and total indexes (S_i^{tot}) according to equations 5 and 8 respectively, requires high computational consumption due to the high number of operations performed to decompose the variance of the model for each parameters and the correlation between them. Fig. 4 shows six graphs, one for each distribution on which the sensitivity analysis was performed. It has to be emphasized that the Sobol sequence [15] (which generates random numbers with low discrepancy) is totally different to the Sobol method for calculating the sensitivity indexes described in section 3.1.

Zhang's method shows that the parameter with greater global sensitivity is *tz*. Furthermore, also the parameter with greater total sensitivity is β. A similar behavior but analyzed from the error propagation perspective in LiDAR-camera calibration is presented in [6]. We can define that the parameters involved directly with the distance of the calibration pattern and image distortions tend to be the most relevant in the error propagation in our calibration system. This error can be minimized in two ways: (1) removing the image distortion and standardizing the images and (2) increasing the reference points in the calibration process. Furthermore, this demonstrates the flexibility to use a calibration approach using a pattern plane.

The second technique implemented for sensitivity analysis of the camera was FAST. This technique gives us the first order indexes S_i by Eq. 14. Table I shows the sensitivity quantified by FAST. It is noted that the parameter *tz* is the more sensitive one. Just as the Sobol method, depth is important. Using a pattern with more corners can reduce this condition. These results confirm experiments made empirically by [16] and also the behaviour shown in Figure 4.



(a) Sobol sequence by MC (b) Halton sequence by MC (c) Gaussian dist. by MC

(d) Gaussian dist. by LHS (e) Uniform dist. by LHS (f) Multivariate dist. by LHS

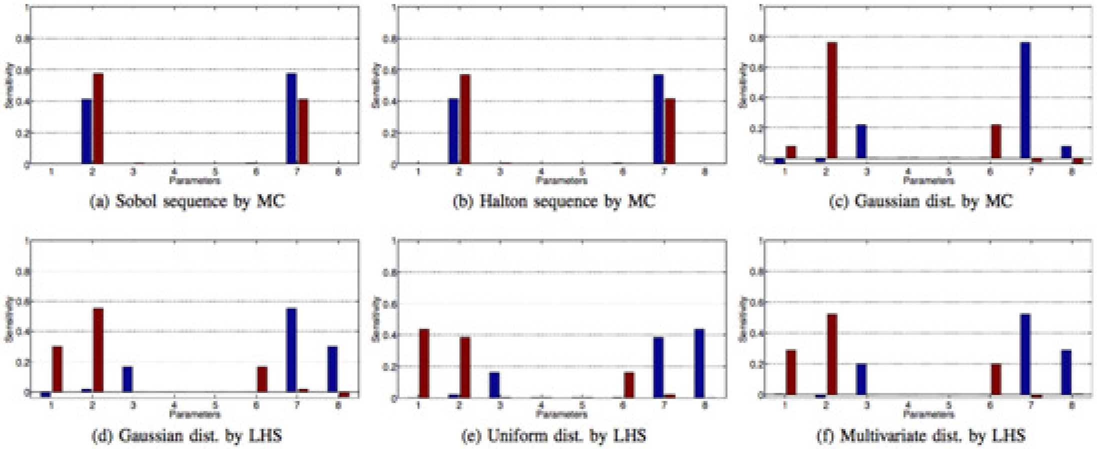
Figure 3 Zhang's calibration method analyzed by Sobol technique, blue bars are the global sensitivity ($\boldsymbol{S_i}$) and red bars are the total sensitivity $\boldsymbol{S_i^{tot}}$**.** Simulated data by MC and LHS with six different distributions. The most relevant parameters are *tz* and $\boldsymbol{\beta}$, i.e. with more sensitive. Table I shows the number parameters reference.

Table 1 Sensitivity analysis by FAST method according to Eq. 14. It is noted that tz and ty parameters are the most sensitive. Only tz parameter agrees with the analysis by the Sobol method, marking a tendency to introduce error in the camera calibration.

Reference	*Parameter*	*Sensitivity index*
1	tx	0.0130
2	ty	0.1108
3	tz	0.5998
4	wx	0.0380
5	wy	0.0329
6	wz	0.0960
7	α	0.0081
8	β	0.0619
9	$u0$	0.0031
10	$v0$	0.0363

4.1.2. LiDAR

The LiDAR sensitivity indexes by Sobol method was calculated as the camera. Figure 5 shows the global (S_i) and total (S_i^{tot}) indexes. It can be seen, that the parameters θ and $\Delta\theta$ are introducing greater uncertainty in the calibration model. The θ parameter corresponds to the LiDAR orientation angle, and is related to the mechanical rotation, while $\Delta\theta$ is a correction of the orientation of each LiDAR laser, because not all of them are in a single plane. Variations in θ and $\Delta\theta$ parameters induce measurement errors because the LiDAR is oriented at a single plane, when actually LiDAR lasers are register a near plane.



(a) Sobol sequence by MC

(b) Halton sequence by MC

(c) Gaussian dist. by MC

(d) Gaussian dist. by LHS

(e) Uniform dist. by LHS

(f) Multivariate dist. by LHS

Figure 4 LiDAR sensitivity analysis by Sobol method, blue bars are the global sensitivity($\boldsymbol{S_i}$) and red bars are the total sensitivity ($\boldsymbol{S_i^{tot}}$). Simulated data by MC and LHS with six different distributions. The most relevant parameters are $\boldsymbol{\theta}$ and $\boldsymbol{\Delta\theta}$, i.e. with more sensitive. Table II shows the number parameters reference

Through the FAST technique, sensitivity indexes match to Sobol technique as shown in Table II. The θ and $\Delta\theta$ are the most sensitive parameters and introducing more noise in the model, and hence increase the output uncertainty error in the calibration values.

Table 2 Sensitivity analysis by FAST technique according to equation 14. It is noted that $\boldsymbol{\theta}$ and $\boldsymbol{\Delta\theta}$ parameters are the most sensitive to introduce error into the laser sensor calibration.

Reference	*Parameter*	*Sensitivity index*
1	ds	0.0140
2	θ	0.4137
3	ϕ	0.0004
4	V_{osc}	0.0160
5	H_{osc}	0.000
6	Δds	0.0006
7	$\Delta\theta$	0.5552
8	$\Delta\phi$	0.000

4.2. Uncertainty

4.2.1. Ladybug2

The uncertainty analysis was conducted to determine the nominal values of the model's error, according to the preliminary sensitivity analysis. Results obtained from a LHS simulation are shown in Table III in which we computed the extrinsic parameters $[R,\boldsymbol{T}]_{\mathrm{W}}^{\mathrm{C}}$ and the intrinsic parameters, image center (u_0, v_0), focal length (α, β).

Table 3 Extrinsic-intrinsic uncertainty camera parameters

	Mean	*Std*
Translation (mm) $\begin{bmatrix} tx \\ ty \\ tz \end{bmatrix}$	$\begin{bmatrix} -125.82 \\ -21.50 \\ 339.43 \end{bmatrix}$	$\begin{bmatrix} 0.0012 \\ 0.0017 \\ 0.0075 \end{bmatrix}$
Rotation (rad) $\begin{bmatrix} roll \\ pitch \\ yaw \end{bmatrix}$	$\begin{bmatrix} -2.22 \\ -2.14 \\ -2.13 \end{bmatrix}$	$\begin{bmatrix} 3.88 \\ 4.37 \\ 4.44 \end{bmatrix} \times 10^{-6}$
Focal (mm)	(551,34,550.81)	(0.0013,0.0014)
Image center (px)	(388.61,506.62)	(0.0025, 0.0029)

Fig. 6 shows graphics simulated uncertainty calculation. Both series of intrinsic and extrinsic parameters exhibit a linear behaviour. Uncertainties were obtained when normal distribution $N(0,\sigma)$ Gaussian noise was added to feature points extracted from the pattern(corners). σ parameter was the result of projecting real-world pattern feature point coordinates' on image plane, its value was 10 pixels. Average error for our projection algorithm was then calculated by measuring the differences between extracted feature points of the whole set of images and those of the real pattern, this error was calculated to a value of *0.25* pixels. Figure 6(a) shows the *Z* axis error tends to increase more than the other axes. Figure 6(c) shows that the error is stable and incremental for both axes while noise increases.

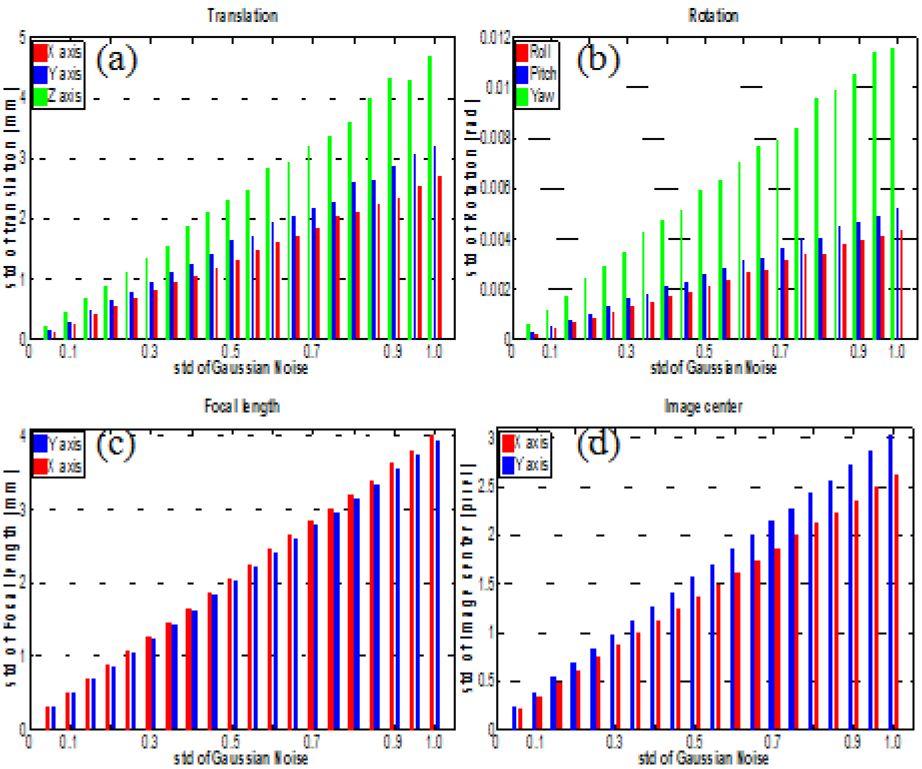


Figure 5 Simulated camera parameters performance. Ordinate axis represents the standard deviation of *2,000* iterations using LHS simulation.

4.2.2. LiDAR

Table IV shows the uncertainty on extrinsic parameters with respect to the world frame. Using a LHS simulation, noise was added to the feature points to evaluate the behaviour of error in these two parameters.

Table 4 Uncertainty on extrinsic LiDAR parameters $[\boldsymbol{R},\boldsymbol{T}]_{\mathrm{W}}^{\mathrm{L}}$

	Mean	*Std*
Translation (mm) $\begin{bmatrix}tx\\ty\\tz\end{bmatrix}$	$\begin{bmatrix}-515.29\\165.62\\-39.06\end{bmatrix}$	$\begin{bmatrix}-1.58\\1.08\\-1.92\end{bmatrix}$
Rotation (rad) $\begin{bmatrix}roll\\pitch\\yaw\end{bmatrix}$	$\begin{bmatrix}0.55\\1.51\\1.52\end{bmatrix}$	$\begin{bmatrix}0.015\\0.011\\0.012\end{bmatrix}$

4. CONCLUSIONS

Variability analysis in the LiDAR-camera calibration is compute. We estimate the relationship between the input and the output quantity in a implicit form. The sensitivity of the calibration depends on the measurements quality, the model used, the calibration method and the conditions under which it is performed. A sensitivity analysis was performed on models of individual calibration of a multi-camera system and a laser sensor. Two techniques are presented for this analysis. Sobol technique, which is based on the decomposition analysis of the variance for each input parameter into the model. And the FAST technique, which uses a Fourier series to represent a multivariable function in the frequency domain using a single variable frequency. For each one of these techniques, we simulated data by two methods: (1) Monte Carlo method and (2) Latin Hypercube sampling method. Since each parameter in the calibration has its own probability distribution, and wanting to generalize the sensitivity analysis, the data were simulated with 6 different types of distributions to cover a higher outlook in this analysis.

It was shown that the parameters tz and β are the most sensitive in the calibration camera model. Sensitivity tests for the calibration method presented in [18] were performed. It was concluded that the parameters involved in the distance between the camera and the calibration pattern (depth) are the most likely to introduce error in the final values of camera calibration, as also concluded in uncertainty analysis in [6] Now that the system behaviour is known, we can pay more attention to characterize the uncertainty in these parameters. LHS method tends to calculate more stable results, i.e. with less variability. On the other hand, it is true that Sobol and FAST techniques allows an extensive sensitivity analysis. The downside is while more parameters the model has more computation time is required to decompose the variance. In our case, the more suitable methodology to simulate data to perform a sensitivity analysis is LHS. In the LiDAR calibration, the parameters θ and $\Delta\theta$ are introducing greater uncertainty in the calibration model. The θ parameter corresponds to the LiDAR orientation angle, and is related to the mechanical rotation, while $\Delta\theta$ is a correction of the orientation of each LiDAR laser. Variations in θ and $\Delta\theta$ parameters induce measurement errors because the LiDAR is oriented at a single plane, when actually LiDAR lasers are registering a near plane.

On the other hand, the uncertainty analysis shows a deviation as small as *0.003 mm* on the translation vector it has been shown that the proposed calibration method is robust and reliable a very small rotation deviation in the order of micro radians has also been obtained. These results contribute to a high level of measurement confidence, despite the complicated working conditions in which our platforms is used. One can now rely on maintaining an acceptable error propagation and uncertainty range in future data fusion operations and 3D texturization. Capture platform has been designed to digitize urban environments, which are mainly a set of planes. For a plane reconstruction (e.g. front of a house) located at 4 meters or less from the system we found a reconstruction error of *2 cm*.

In future works, using some other calibration methods, we expect to be able to minimize this error propagation and uncertainty. Mainly by using data from different calibration instruments. At the end we will be able of compensating the error in 3D reconstruction.

REFERENCES

[1] GEB Archer, A. Saltelli, and IM Sobol. Sensitivity measures, anova-like techniques and the use of bootstrap. Journal of Statistical Computation and Simulation, 58(2):99–120, 1997.

[2] Dan G Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. Sensitivity and Uncertainty Analysis: Applications to large-scale systems, volume 2. CRC Press, 2004.

[3] Roger A Cropp and Roger D Braddock. The new morris method: An efficient second-order screening method. Reliability Engineering & System Safety, 78(1):77–83, 2002.

[4] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. Statistics and computing, 10(3):197–208, 2000.

[5] Angel-Iván García-Moreno, José-JoelGonzalez-Barbosa, Francisco-Javier Ornelas-Rodriguez, Juan B Hurtado-Ramos, and Marco-Neri Primo-Fuentes. Lidar and panoramic camera extrinsic calibration approach using a pattern plane. In Pattern Recognition. Springer, 2013.

[6] Angel-Iván García-Moreno, Denis-Eduardo Hernandez-García, José-Joel Gonzalez-Barbosa, Alfonso Ramírez-Pedraza, Juan B Hurtado-Ramos, and Francisco-Javier Ornelas-Rodriguez. Error propagation and uncertainty analysis between 3d laser scanner and camera. Robotics and Autonomous Systems, 62(6):782–793, 2014.

[7] J.C. Helton, F.J. Davis, and J.D. Johnson. A comparison of uncertainty and sensitivity analysis results obtained with random and latin hypercube sampling. Reliability Engineering and System Safety, 89(3):305–330, 2005.

[8] L. Lilburne and S. Tarantola. Sensitivity analysis of spatial models. International Journal of Geographical Information Science, 23(2):151–168, 2009.

[9] Hervé Monod, Cédric Naud, and David Makowski. Uncertainty and sensitivity analysis for crop models. D. WALLACH, D. MAKOWSKI et J. JONES, éditeurs: Working with Dynamic Crop Models, pages 55–100, 2006.

[10] Gustavo G Pilger, Joao Felipe CL Costa, and Jair C Koppe. Improving the efficiency of the sequential simulation algorithm using latin hypercube sampling. In Geostatistics Banff 2004. Springer, 2005.

[11] Myers RH. Response surface methodology – current status and future directions. J Qual Technol, 31(1):30–44, 1999.

[12] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer Physics Communications, 181(2):259–270, 2010.

[13] Andrea Saltelli, Karen Chan, E Marian Scott, et al. Sensitivity analysis. Wiley New York, 2000.

[14] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons, 2004.

[15] I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. {USSR} Computational Mathematics and Mathematical Physics, 7(4):86–112, 1967.

[16] Wei Sun and Jeremy R Cooperstock. An empirical evaluation of factors influencing camera calibration accuracy using three publicly available techniques. Machine Vision and Applications, 17(1):51–67, 2006. [17]Stefan J Wijnholds and A-J Van Der Veen. Multisource self-calibration for sensor arrays. Signal Processing, IEEE Transactions on, 57(9):3512–3522, 2009.

[18] Z. Zhang. A flexible new technique for camera calibration. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1330–1334, 2000.

AUTHORS

Ángel Iván García Moreno. He is currently Ph.D. candidate in Advanced technology from the Research Center for Applied Science and Advanced Technology. He obtained his M.Sc. degree from the National Polytechnic Institute, in 2012, in the computer vision area. He received the B.S. degree in Informatics by the Autonomous University of Queretaro in 2009. His personal interests are computer vision, remote

José Joel González Barbosa was born in Guanajuato, Mexico, in 1974. He received the M.S. degree in Electrical Engineering from the University of Guanajuato, Mexico, and Ph.D. degree in Computer Science and Telecommunications from National Polytechnic Institute of Toulouse, France, in 1998 and 2004, respectively. He is an Associate Professor at the CICATA Querétaro-IPN, Mexico, where he teaches courses in Computer Vision, Image Processing and Pattern Classification. His current research interests include perception and mobile robotics.

Juan B. Hurtado Ramos. He received the B.S. degree in Communications and Electronics Engineering in 1989 by the Guadalajara University. He obtained his Ph.D. degree from the Optics Research Center of the University of Guanajuato in 1999. He is currently a Profesor-Researcher of the CICATA Querétaro-IPN, Mexico in the Image Analysis group. His personal interests are mainly in the field of Metrology using optical techniques. Since 1998, he is member of the Mexican National Researchers System.

Francisco Javier Ornelas Rodriguez. He received the B.S. degree in Electronic Engineering in 1993 by University of Guanajuato. He obtained his Ph.D. degree from the Optics Research Center in 1999. He is currently a Profesor-Researcher of CICATA Querétaro-IPN, Mexico in the Image Analysis team. His personal interests are mainly in the field of Metrology using optical techniques. He is member of the Mexican National Researchers System.

INTENTIONAL BLANK

IMAGE SEARCH USING SIMILARITY MEASURES BASED ON CIRCULAR SECTORS

Jan Masek, Radim Burget, Lukas Povoda and Martin Harvanek

Department of Telecommunications, Faculty of Electrical Engineering,
Brno University of Technology, Brno, Czech Republic
masek.jan@phd.feec.vutbr.cz, burgetrm@feec.vutbr.cz,
xpovod00@stud.feec.vutbr.cz, xharva01@stud.feec.vutbr.cz

ABSTRACT

With growing number of stored image data, image search and image similarity problem become more and more important. The answer can be solved by Content-Based Image Retrieval systems. This paper deals with an image search using similarity measures based on circular sectors method. The method is inspired by human eye functionality. The main contribution of the paper is a modified method that increases accuracy for about 8% in comparison with original approach. Here proposed method has used HSB colour model and median function for feature extraction. The original approach uses RGB colour model with mean function. Implemented method was validated on 10 image categories where overall average precision was 67%.

KEYWORDS

CBIR, circular sectors, cross-validation, image features, image processing, image similarity, optimization

1. INTRODUCTION

Nowadays, the amount of transmitted image data through internet is every day still growing and due to this fact digital image databases are filled with new terabytes of images. In order to search and manage this data, there is strong need to index or categorize these images using proper system. Searching images on the basis of similarity can be used in medicine, arts, industry [1], security, military and many other areas [2].

This work deals with an image categorization and search on the basis of content. Systems that provide this functionality are called Content-Based Image Retrieval (CBIR) [3]. These systems search huge image databases, where for every image the special signature is created. The signature is used for comparing with image we want to categorize. In our approach we improved circular sector method introduced in [4] and we increased accuracy for about 8%.

CBIR systems usually use visual image properties like colour, texture and shape for creating feature vectors that are saved in to the database. Visual image properties are compared by using similarity measurements (Euclidean metrics, Manhattan metrics) and according to the value of

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 241–251, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51519

measurements, images are compared or searched in database. CBIR systems use several methods for the computing of feature vectors. Methods can be based on local or global feature extraction or can be based on colour coherence vectors [5], colour moments [6], circular sectors [4] or Gabor filters [6]. The CBIR system architecture is depicted in Figure. 1.

The main contribution of the paper is method that modifies original approach [4]. This approach uses circular sectors method that is inspired by human eye functionality. We achieved higher accuracy for about 8% when compared with [4]. We conducted parameter optimizations using cross validation process and machine learning [19] to find optimal learning algorithm and its configuration. Our approach uses different types of circular sector features where we used HSB colour model with median function instead of RGB colour model with mean function for feature computation.

The rest of this paper is organized as follows: The second section describes related work with focus on CBIR systems. Section 3 describes circular sector method. In section 4 method modification is described. Image data sets are described in section 5. The section 6 describes optimization of parameters. Results are discussed in section 7 and section 8 concludes this paper.

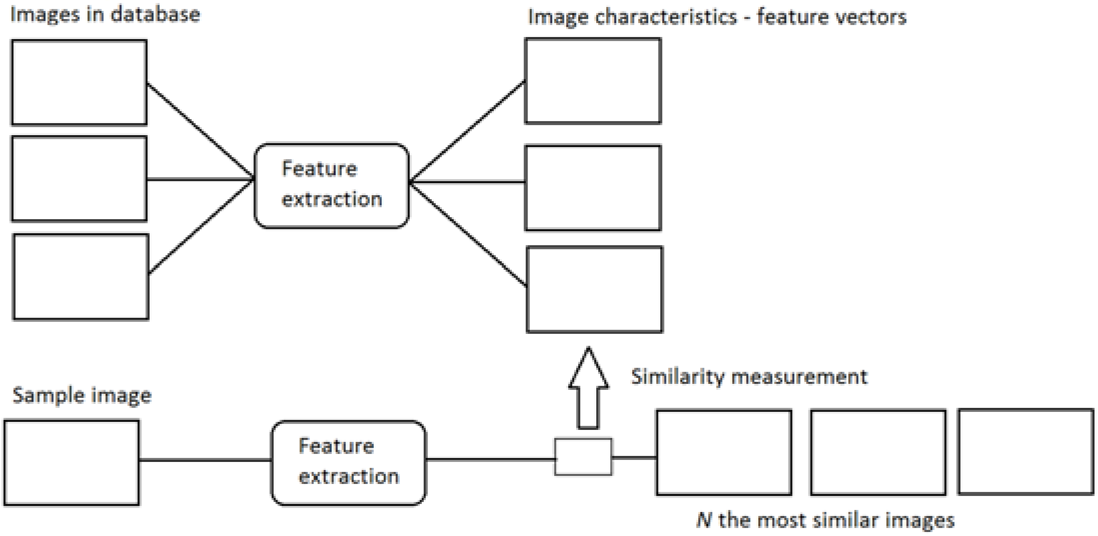


Figure 1. Content base image retrieval system architecture.

2. RELATED WORK

Until today many content base image retrieval systems have been created [3]. We present several leading systems in this chapter. For example QBIC system from IBM has been used for many further work dealing with CBIR. Another leading systems are visualSeek or Netra [4]. From these systems many following system have been derived [7], [8] and [9].

There are many works dealing with different image features. Histogram intersection computation has been used to compare images in [10]. Cumulative histograms were described in [11] and spatial matching with colour histograms were described in [12]. In [13] and [14] is proven that colour features are very suitable for similarity measurements.

We also described method based on dominant colours in [15] for measuring image similarity and in [16] system for automatic image labelling using similarity measures is described. In [17] video scenes were segmented using similarity measures.

3. Circular Sectors Method

This method has been described in [4] and it is based on human eye principle. The human eye firstly focuses on the center of image and then goes to the edges of image. The method creates special image features that are obtained from image. Firstly, the center of image is determined and then image is divided in to concentric circles *Ci*, where *i* is number of circle. Then every circle is divided to sectors *Si,* where *Si* = 8 *Ci*. In this case, seven circles are chosen and 252 sectors in whole image are created (see Figure 2 - left). Due to the fact that we use RGB color model with 3 channels, we need to create 3·252 = 756 sectors. The next step is to compute mean value in every sector (see Figure 2 - right). So the image feature is computed as mean value of defined sector. When this method is applied on input image, output feature vector containing mean values of all sectors is computed.

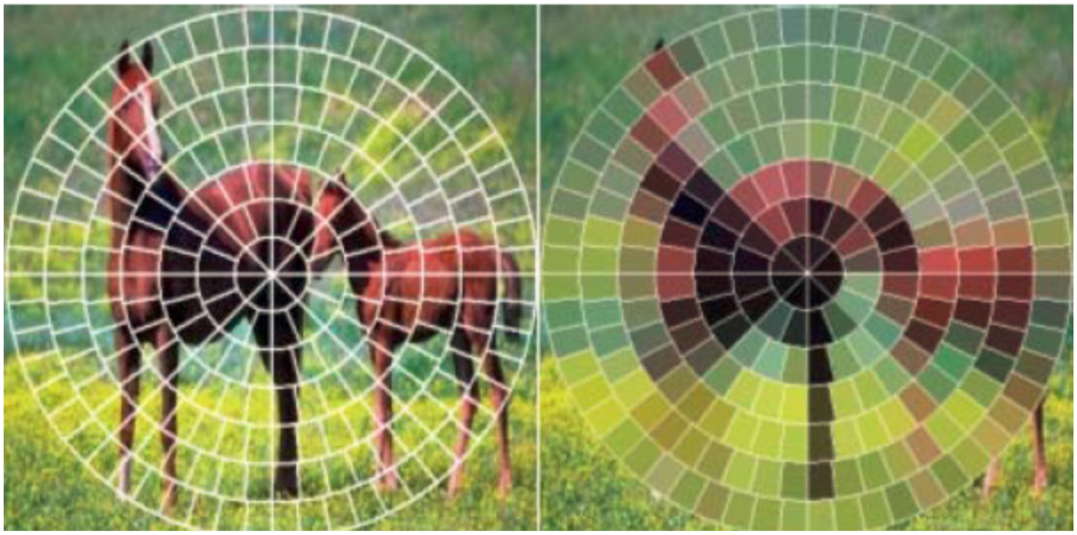


Figure 2. Circular sectors in the image (left), average colour values in each sector (right) [4].

To make this method rotation-invariant, the mean values of sectors are sorted in every circle. Figure 3 shows that sorted sectors are similar when using normal or rotated image.

4. Method Modification

In originally described method authors used for feature extraction RGB channels and mean value computation. We decided to create new image features using median function and using HSB colour model. The comparison between these new features and previously used features will be described in chapter 6. Our implementation of algorithm has been created in JAVA programming language according to previous work [4] with our new modifications.

4.1 Median

The mean colour value of sector cannot exactly determine the distribution of pixel values (e.g. if image contains a little noise). We modify previous method with using median values instead of mean values.

4.2 HSB Colour Model

For human perception the HSB model suits better than RGB model. HSB is an abbreviation of Hue, Saturation, and Brightness. This model use the cylindrical-coordinate representations of values in an RGB model. We use the HSB model instead of RGB colour model.

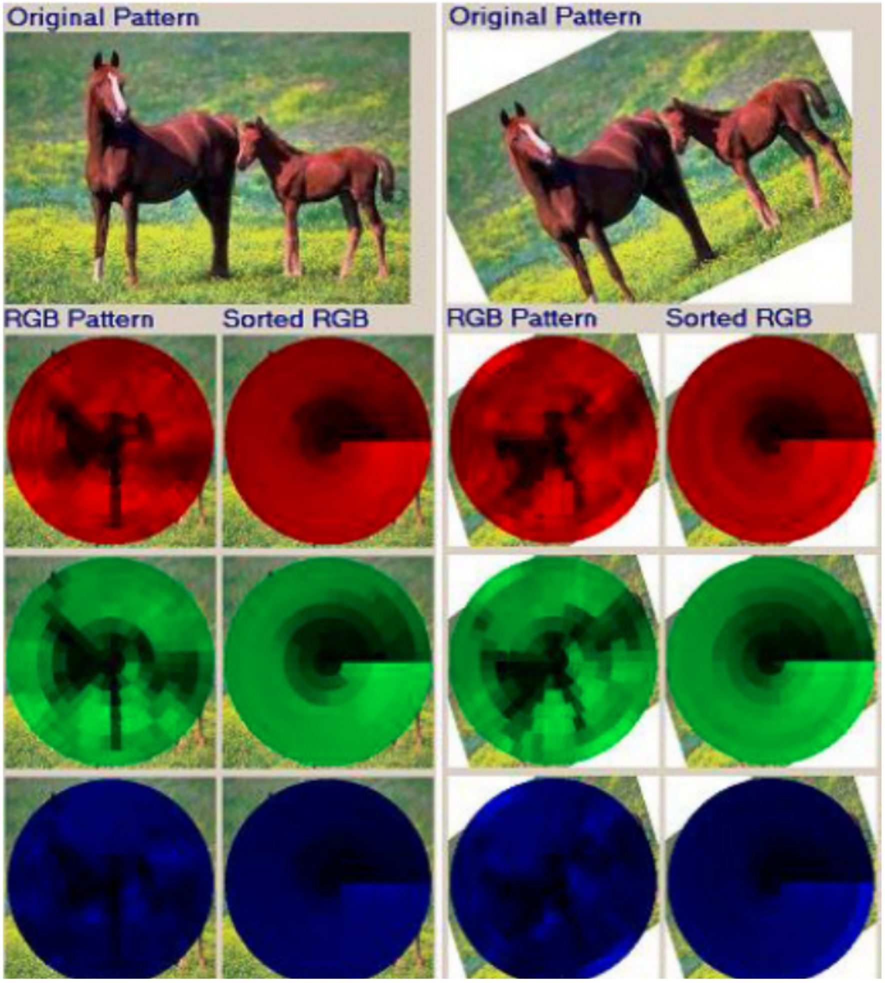


Figure 3. Original image with original and sorted sector values (left), 30° rotated image with original and sorted sector values [4].

5. Data Sets

In this work we used same image data set as authors that described original circular sector method [4]. This data set is available to download from [20]. Data set consists of 10 categories (ancient, beach, bus, dinosaur, elephant, flower, food, horse, mountain, natives) where every category contains 100 images. We have 1000 images overall. The images have dimension 354x256 pixels. The example of used images is shown in Figure 4.

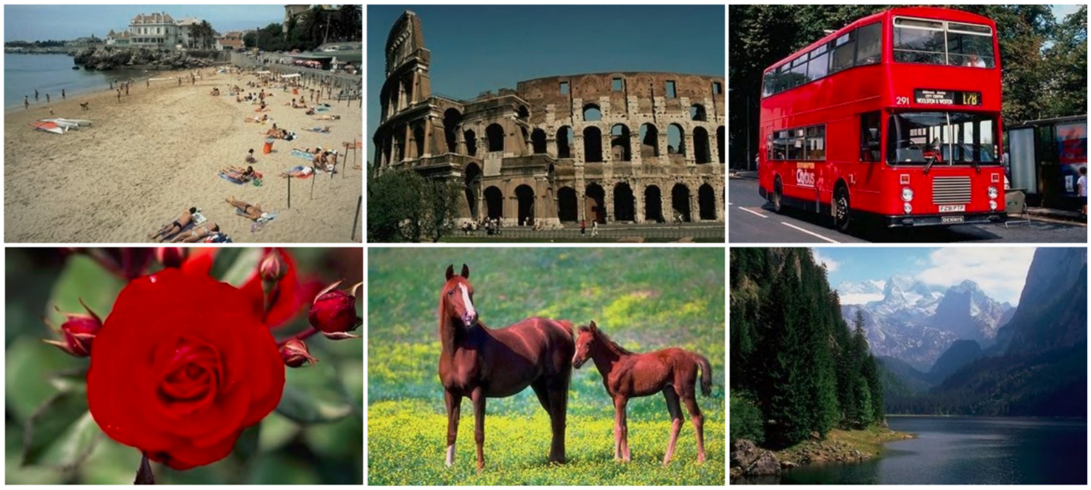


Figure 4. Example of used images

6. Parameters Optimization

There are many options how to extract features from image. For example dimensions of image, the number of circles for creating sectors. Features can also be extracted using RGB or HSB colour model or computing median or mean value. We chose nine variants that we wanted to compare. For every variant, features were generated to format suitable for RapidMiner [21] data mining tool. This tool contains many machine learning algorithms (e.g. algorithms of artificial intelligence, optimization algorithms). We used cross-validation process [19] (see Figure 5) that computes accuracy for every variant. The cross validation process used SVM (Support Vector Machines) algorithm [18] of artificial intelligence. The SVM algorithm had these parameters:

- SVM type: C-SVC
- Kernel type: linear
- C: 1.1
- Epsilon: 0.001

The results of cross-validation process for every variant is shown in Table 1. It shows that HSB colour model has higher accuracy than RGB model and also median function achieves higher accuracy than mean function. The best achieved accuracy is 75.6% for image with 400x400 pixels dimensions, with 7 circles and HSB model where features are computed using median function. It also shows that our approach that uses HSB model with median function has higher accuracy (75.6%) in comparison with original approach [4] that uses RGB model with mean function. Our modified method achieves for about 8% higher accuracy.

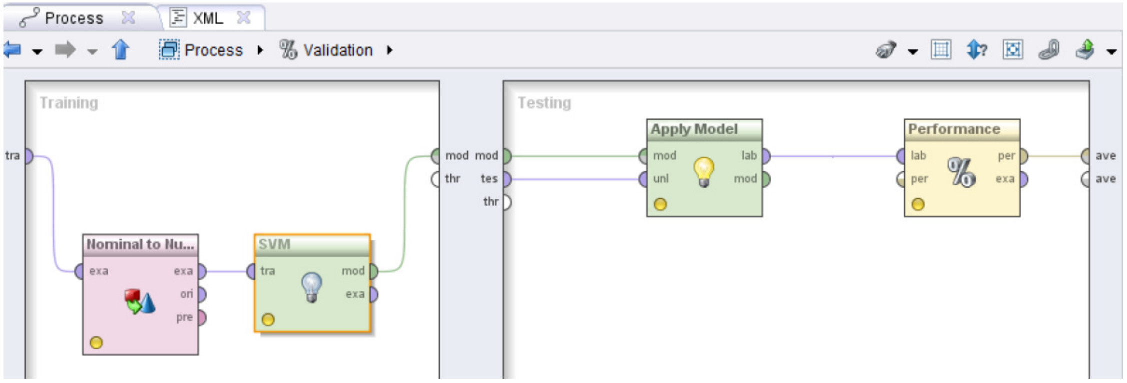


Figure 1. The scheme of cross validation process in RapidMiner tool.

Table 1. Selected variants and their accuracy of classification.

Dimensions	Circles	RGB mean	RGB median	HSB mean	HSB median
200x200	3	64.3 %	66.0 %	71.0 %	71.5 %
200x200	5	65.6 %	67.3 %	72.3 %	72.9 %
200x200	7	68.1 %	70.5 %	72.6 %	74.8 %
300x300	3	65.0 %	65.6 %	70.6 %	72.4 %
300x300	5	68.6 %	68.9 %	72.8 %	72.0 %
300x300	7	67.2 %	71.3 %	72.9 %	74.1 %
400x400	3	64.4 %	65.5 %	71.1 %	72.2 %
400x400	5	68.6 %	69.0 %	72.7 %	72.4 %
400x400	7	67.6 %	70.8 %	73.2 %	**75.6 %**

Table 2 shows confusion matrix for every image category. The best precision was achieved with dinosaur category (97.09%) and the lowest precision was achieved category ancient (53.45%).

Table 2. Confusion matrix for parameters (dimensions 400x400, circles 7, HSB median).

		Label (real values)										**Prec. [%]**
		ancient	beach	bus	dinosaur	elephant	flower	food	horse	mountain	natives	
Prediction	ancient	62	16	0	0	7	0	3	2	12	14	53.45
	beach	11	62	3	0	1	0	5	2	20	2	58.49
	bus	2	2	83	0	0	3	3	0	5	1	83.84
	dinosaur	0	1	0	100	0	0	1	0	0	1	**97.09**
	elephant	8	2	0	0	79	0	1	0	3	6	79.80
	flower	0	0	4	0	0	88	4	0	0	2	89.80
	food	0	3	4	0	0	7	72	1	2	11	72.00
	horse	3	1	0	0	1	0	2	93	0	0	93.00
	mountain	6	12	4	0	6	2	1	0	57	3	62.64
	natives	8	1	2	0	6	0	8	2	1	60	68.18

7. RESULTS

We performed several comparison tests to verify our modified method. For evaluation, we used precision that is computed:

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}$$

where N_{TP} is a number of true positive (relevant) images and N_{FP} is number of false positive (irrelevant) images. Firstly, one pattern image is selected and its feature vector is computed, then this feature vector is compared with the feature vectors of all images from data set. When data set contains 1000 images, the comparison process had to be executed 1000000 times.

Comparison has been done with computing Euclidean and Manhattan metrics

$$d_E(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

$$d_M(x,y) = \sum_{i=1}^{d}|x_i - y_i|$$

where d is the length of input feature vector and x and y are feature vectors of 2 images that are being compared. For every image, N the most similar images are selected, where we set $N = \{10,\ 25, 50, 100\}$ and the precision is computed for N images. Finally, the overall precision is computed as average of all precisions computed for every image.

Table 3 shows precision of every category (each contains 100 images) using Euclidean metrics and Table 4 shows precision using Manhattan metrics. Overall average precision is shown in Table 5. The best achived precision was 67.23% for $N = 10$ with using Manhattan metrics.

Table 3. Precision of every category using Euclidean metrics.

N	Ancient [%]	Mountain [%]	Bus [%]	Dinosaur [%]	Elephant [%]	Food [%]	Horse [%]	Beach [%]	Flowers [%]	Natives [%]
10	44.3	51.7	62.5	**97.7**	65.2	56.9	89.8	53.9	72.2	43.3
25	33.48	44.44	52	97.28	52.48	44.68	82.24	44.04	59.16	33.92
50	30.004	40.58	42.4	94.8	44.72	36.54	73.04	37.56	43.96	28.52
100	25.73	34.26	34	80.44	37.76	28.79	56.24	31.51	29.95	24.81

Table 4. Precision of every category using Manhattan metrics.

N	Ancient [%]	Mountain [%]	Bus [%]	Dinosaur [%]	Elephant [%]	Food [%]	Horse [%]	Beach [%]	Flowers [%]	Natives [%]
10	51	52	63.9	**99.7**	66	66	92.1	50.3	79.2	52.1
25	41.76	44.72	54.24	99.52	53.64	58	86.96	43.04	67.8	43.88
50	33.76	39.66	45.22	98.84	44.74	48.54	79.18	37.04	51.82	37.8
100	28.59	34.04	36.63	91.81	36.98	37.01	62.42	32.28	36.55	32.46

Table 5. Overall average precision

N	Euclidean distance	Manhattan distance
10	63.75 %	**67.23 %**
25	54.57 %	59.36 %
50	47.22 %	51.66 %
100	38.25 %	42.74 %

All computations were performed on computer with processor Intel Core i5 2.5 GHz and with 4GB of RAM memory. The computing of feature vector for all images took 1 minute and 9 seconds. To find and compare input pattern image with all image feature vectors (1000) took approximately 2 seconds.

The results of searching pattern image (see Figure 6) for horse category are shown in Figure 7. When pattern image is rotated to left by 90°, the results (see Figure 8) contain 4 incorrectly selected images.

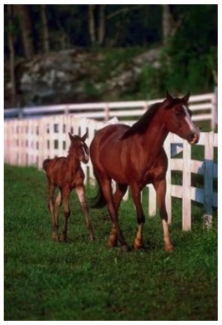


Figure 6. Pattern image for horse category.

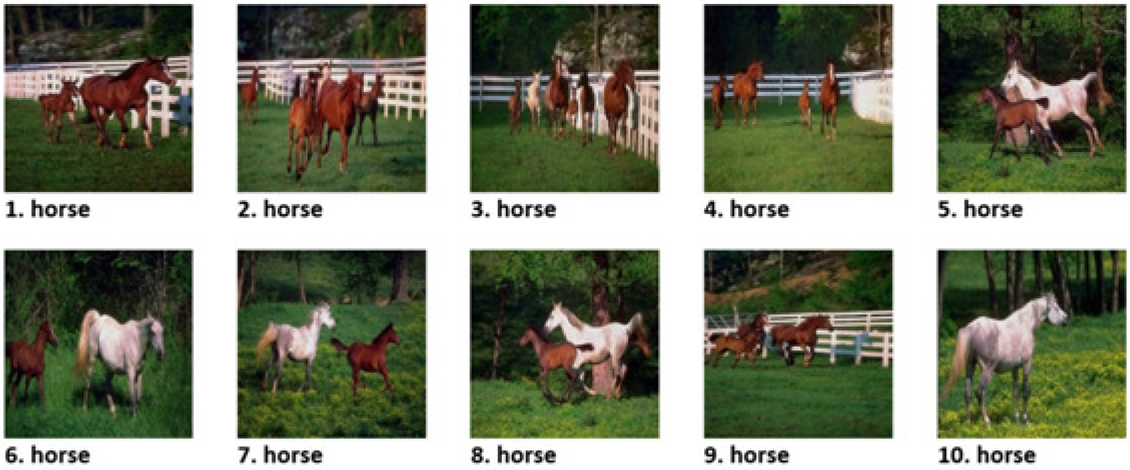


Figure 7. First 10 the most similar images of horse pattern image.

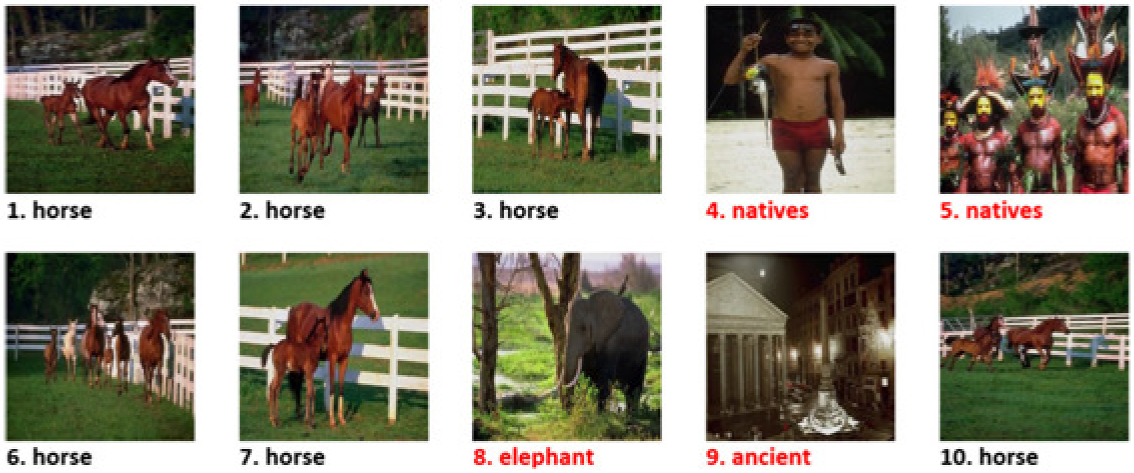


Figure 8. First 10 the most similar images of horse pattern image rotated about 90°

8. CONCLUSION

The main contribution of this paper is a method that increases accuracy in CBIR systems for about 8% in comparison with original approach [4]. The origin achieved accuracy was 67.6%. We are currently able to achieve 75.6% accuracy with using the same image data set. We tried to find suitable parameters for circular sectors method. We selected the method because it is inspired by human eye functionality. We conducted parameters optimization using cross validation process with algorithms of artificial intelligence, where we found that HSB colour model and median function for feature computation achieve better result than original approach using RGB colour model with mean function for feature computation. For testing we used 1000 images from 10 categories. The best result of average precision was 67.23% with using Manhattan metrics. The average time for image comparison with database was 2 seconds on common computer.

ACKNOWLEDGEMENTS

Research described in this paper was financed by the National Sustainability Program under grant LO1401 and by the Czech Science Foundation under grant no. 102/12/1274 and with MPO FR-TI4/151, Czech Republic. For the research, infrastructure of the SIX Center was used.

REFERENCES

[1] A. Rangkuti, Haris, et al. Analysis of Image Similarity with CBIR Concept Using Wavelet Transform and Threshold Algorithm. In: Computers & Informatics (ISCI), 2013 IEEE Symposium on. IEEE, 2013. pp. 122-127.

[2] E. Chalom, Asa Eran, Biton, Elior. Measuring image similarity: an overview of some useful applications. IEEE Instrumentation & Measurement Magazine, IEEE, 2013, vol. 16, no. 1, pp. 24-28.

[3] F. Long, H. Zhang and D. Dagan Feng., "Fundamentals of Content-Based Image Retrieval". Multimedia Information Retrieval and Management–Technological Fundamentals and Applications, Springer-Verlag, pp. 1-26, 2003.

[4] Omar, Samia G., Ismail, Mohamed A.; Ghanem, Sahar M. WAY-LOOK4: A CBIR system based on class signature of the images' color and texture features. In: Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS, International Conference on. IEEE, 2009. pp. 464-471

[5] Pass, Greg, Zabih, Ramin, Miller, Justin. Comparing images using color coherence vectors. In: Proceedings of the fourth ACM international konference on Multimedia. ACM, 1997. pp. 65-73.

[6] Singh, S. Mangijao, Hemachandran, K Content-Based Image Retrieval using Color Moment and Gabor Texture Feature International Journal of Computer Science Issues (IJCSI), 2012, vol.9, no.5, s. 299-309.

[7] Y. Rui and T. S. Huang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues", Journal of Visual Communication and Image Representation 10, pp. 39–62, 1999.

[8] T. Wang, Y. Rui, J. Guang, Sun, "Constraint Based Region Matching for Image Retrieval", International Journal of computer vision 56 1/2, pp. 37-45, 2004.

[9] A. M. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1380, 2000.

[10] Michael J. Swain and Dana H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11–32, June 1991.

[11] Markus Stricker and Markus Orengo, "Similarity of color images," in In Proceedings of SPIE Storage and Retrieval for Image and Video Databases, 1995, pp. 381–392.

[12] Markus Stricker, Alexander Dimai, and Er Dimai, "Color indexing with weak spatial constraints," in In Proceedings of SPIE Storage and Retrieval for Image and Video Databases, 1996, pp. 29–40.

[13] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, "An online algorithm for large scale image similarity learning," in Advances in Neural Information Processing Systems, 2009.

[14] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, "Large scale online learning of image similarity through ranking," Journal of Machine Learning Research, vol. 11, pp. 1109–1035, March 2010.

[15] J. Karasek, R. Burget, V. Uher, J. Masek, M. Dutta, Color Image (Dis) Similarity Assessment and Grouping based on Dominant Colors. In 2014 37th International Conference on Telecommunications and Signal Processing (TSP).Berlin, Germany: 2014. pp. 631-634. ISBN: 978-80-214-4983- 1.

[16] V. Uher, R. Burget, J. Karasek, J. Masek, M. Dutta, M. Automatic Image Labelling using Similarity Measures. InMEDCOM 2014 CD-ROM. Greater Noida: IEEE, 2014. pp. 101-104. ISBN: 978-1-4799-5096- 6.

[17] R. Burget, K. Ray, V. Uher, J. Masek, M. Dutta, Supervised Video Scene Segmentation using Similarity Measures Supervised Video Scene Segmentation using Similarity Measures. In 36th International Conference on Telecommunications and Signal processing. 2013. pp. 793-797. ISBN: 978-1-4799-0402- 0.

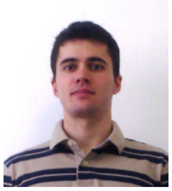
[18] Chang, Chih-Chung, Lin, Chih-Jen. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, vol. 2, no. 3, s. 27.

[19] Akthar, Fareed, Hahne, Caroline. RapidMiner 5: Operator Reference. Dortmund: Rapid-I GmbH, 2012. 990 s

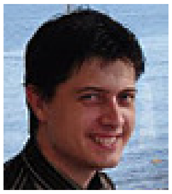
[20] http://wang.ist.psu.edu/docs/related.shtml

[21] https://rapidminer.com/

AUTHORS



Jan Masek is Ph.D. student at the Department of Telecommunications, Faculty of Electrical Engineering, Brno University of Technology, Brno, Czech Republic. He obtained his MSc. in 2012 (Communications and Informatics). He is interested in image processing, data mining, parallel systems.



Dr. Radim Burget is associated professor at the Department of Telecommunications, Faculty of Electrical Engineering, Brno University of Technology, Brno, Czech Republic. He obtained his MSc. in 2006 (Information Systems) and his finished his Ph.D. in 2010. He is associated professor since 2014. He is interested in image processing, data mining, genetic programming and optimization.

Lukas Povoda is Ph.D. student at the Department of Telecommunications, Faculty of Electrical Engineering, Brno University of Technology, Brno, Czech Republic. He obtained his MSc. in 2014 (Communications and Informatics). He is interested in image processing, text processing, and genetic programming.

Martin Harvanek obtained his MSc. in 2014 at the Department of Telecommunications, Faculty of Electrical Engineering, Brno University of Technology, Brno, Czech Republic. He is interested in image processing and data mining.

INTENTIONAL BLANK

IMPLEMENTATION OF A BPSK MODULATION BASED COGNITIVE RADIO SYSTEM USING THE ENERGY DETECTION TECHNIQUE

Sid Ahmed Chouakri[1], Mohammed El Amine Slamat[1,*]
and Abdelmalik Taleb-Ahmed[2]

[1]Laboratoire de Télécommunications et Traitement Numérique du Signal,
University of Sidi Bel Abbes, ALGERIA
sa_chouakri@hotmail.com
* aimenslamat20@gmail.com
[2]Laboratoire LAMIH, Université de Valenciennes et du Hainaut Cambrésis,
FRANCE
taleb@univ-valenciennes.fr

ABSTRACT

We present in this work an energy detection algorithm, based on spectral power estimation, in the context of cognitive radio. The algorithm is based on the Neyman-Pearson test where the robustness of the appropriate spectral bands identification, is based, at one hand, on the 'judicious' choice of the probability of detection (P_D) and false alarm probability (P_F). First, we accomplish a comparative study between two techniques for estimation of PSD (Power Spectral Density): the periodogram and Welch methods. Also, the interest is focused on the choice of the optimal duration of observation where we can state that this latter one should be inversely proportional to the level of the SNR of the transmitted signal to be sensed. The developed algorithm is applied in the context of cognitive radio. The algorithm aims to identify the free spectral bands representing, reserved for the primary user, of the signal carrying information, issued from an ASCII encoding alphanumeric message and utilizing the BPSK modulation, transmitted through an AWGN (Added White Gaussian Noise) channel. The algorithm succeeds in identifying the free spectral bands even for low SNR levels (e.g. to -2 dB) and allocate them to the informative signal representing the secondary user.

KEYWORDS

Radio cognitive, energy detection, spectrum sensing, power spectral density, BPSK modulation, primary/secondary user

1. INTRODUCTION

It is widely recognized that wireless digital communications systems do not exploit the entire available frequency band. Future wireless generations' systems will therefore have to take advantage of the existence of such unoccupied frequency bands, thanks to their ability to listen and adapt to their environment [1]. The recent rapid evolution of wireless leads a strong demand in terms of spectrum resources. To overcome this problem it must be a good spectrum management and therefore a more efficient use of it [2-4]. The recent researches show that, 80%

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 253–263, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51520

to 85% of the total spectrum is unused, while only 15% to 20% of the spectrum is used for the maximum period of time [5-6].

Cognitive radio has emerged as a key technology, which allows opportunistic spectrum access and respond directly to the needs related to the management of the environment of the radio terminal [7-8]. Cognitive radio (CR) is basically software-defined radio SDR with artificial intelligence, able to sense and react to their changing environment [9-11]. In 1998, at the royal Institute of technology KTH, Joseph Mitola III exhibited his work on a radio that is aware of the electromagnetic environment, which is able to change the behavior of its physical layer and which can adopt complex strategies. Cognitive radio (CR) is the name of this new approach to communication in wireless networks [12].

The paper is organized as follows: In the second section of this work we will present the basic concepts of cognitive radio, as well as its main features. In the third section we will introduce a so important topic in the radio cognitive RC system which is the technique for the detection of 'available' spectrum band that is the energy detection technique based on spectral power estimation.

In the fourth section, we will discuss two main statistical tests, upon which is based the energy detection technique, that are the Bayesian test and the Neyman-Pearson test. The latter one is based on the calculation of the probability of false alarm P_F and the probability of detection P_D. In the fifth and last section we will introduce the technique of detection of energy in the context of a cognitive radio scenario, articulated on the QPSK modulation, and a procedure that allows inserting the secondary user in the unoccupied band.

2. OVERVIEW ON COGNITIVE RADIO

2.1. Architecture of Cognitive Radio Networks

A detailed description of the architecture of the cognitive radio networks is vital to develop effective communication protocols. The elements that compose the CRN cognitive radio networks are represented in figure 1 [13].The architecture of the cognitive radio systems is articulated upon two distinct networks: primary and secondary [14]. The primary network is licensed to use certain spectral bands. The primary network acquired that right through the purchase of licenses from government agencies, e.g. cellular networks, the broadcast TV networks, etc. The secondary network (known as cognitive radio, dynamic access networks, or unlicensed network) is a network that has no license to operate on the spectral band. However, thanks to the additional features they have, these users can share the spectral bandwidth with the primary users provided they do not harm their transmissions or take advantage of their absence to transmit.

2.2. Functions of Cognitive Radio

The main functions of cognitive radios are:

2.2.1. Spectrum sensing:

It is a fundamental function allowing the cognitive radio users to detect the spectrum used by primary systems and improve the efficiency of the total spectrum [5].

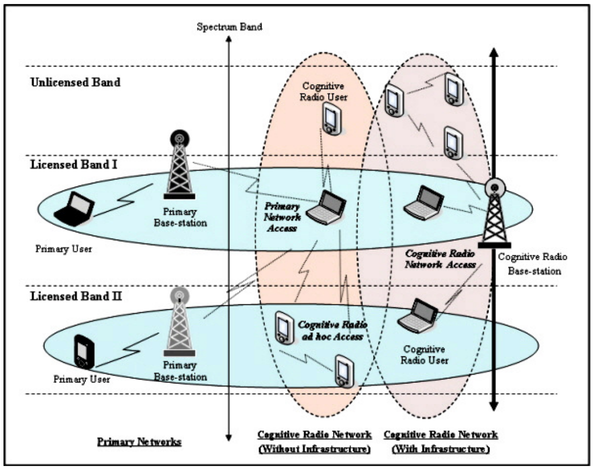


Figure 1. Coexistence between two network types: primary and secondary network [13].

2.2.2. Spectrum management:

The unused spectral bands have different characteristics from the others. All of this information changes over time given the dynamic nature of the radio environment.

2.2.3. Spectrum mobility:

The definition of mobility of the spectrum is to maintain the requirement of communication seamless during the transition to a better spectrum.

3. The Spectrum Sensing Based Energy Detection

Figure 2 represents the details of the classification of the spectrum sensing techniques.

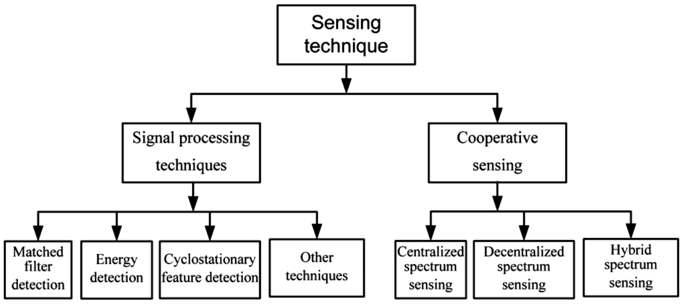

Sensing technique
Signal processing techniques
Cooperative sensing
Matched filter detection
Energy detection
Cyclostationary feature detection
Other techniques
Centralized spectrum sensing
Decentralized spectrum sensing
Hybrid spectrum sensing

Figure 2. Classification of the spectrum sensing techniques [15]

3.1. Mathematical Basis

When information about the presence of the Gaussian noise is available, the energy detection approach is a suitable technique for spectrum sensing. Receivers do not need an exhaustive knowledge of primary users. The energy detection (ED) simply deals with the primary signal as noise and decides the presence or absence of the primary signal based on the energy of the observed signal [5]. This technique measures the received energy of primary user. If the energy is less than a certain threshold value then it decides as free band. Figure 3 illustrates the energy detection technique methodology block diagram [16]. The band pass filter BPF selects the center frequency and the bandwidth of interest. The filter is followed by a squared rising to measure the energy of the received signal. Subsequently, it is the integration phase. Finally the integrator output is compared with a threshold to decide if the primary user is present or not.

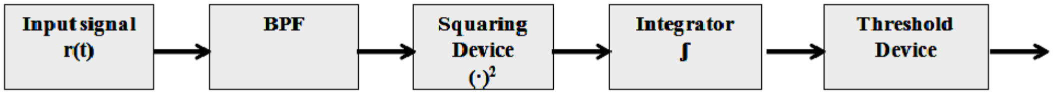


Figure 3. Energy detection technique methodology block diagram [16].

The sample of the signal received by the secondary user can be represented by [5-6]:

$$\begin{aligned} H_0: &\quad y(n)= w(n) \\ H_1: &\quad y(n)=s(n)+w(n) \end{aligned} \qquad (1)$$

where s(n) is the signal to be detected, w(n) is the added white Gaussian noise (AWGN), and n is the index of sample. H_0 is the hypothesis that the primary user is absent and H_1 represents the hypothesis that the primary user is present. Metric decision for the detection of energy may be written by [5-6]:

$$\mathrm{T}= \sum_{i=1}^{N} y_i{}^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \qquad (2)$$

where N is the dimension of the observation vector. The decision on the occupation of a band can be obtained by comparing the T metric decision against a threshold λ.

The performance of the detection algorithm can be summarized by two probabilities: the detection probability P_D and false alarm probability P_F. It can be formulated as [2]:

$$P_D=P_r(T>\lambda\backslash H_1) \qquad (3)$$

P_F is the probability that the test decides incorrectly that the reporting frequency is occupied when actually is not, and it can be written by:

$$P_F=P_r(T>\lambda\backslash H_0) \qquad (4)$$

3.2. The Calculation of the Threshold

Let us assume the model of the received signal given by equation (1); where s (n) is with zero mean value and a variance of $\sigma_s{}^2$, w (n) is with zero mean value and a variance of $\sigma_w{}^2$, and n = 1, 2, 3... N is the observation sample.

Let us suppose: *cov* $(w_i, w_j)= 0 \ \forall\, i \neq j$, *cov* $(s_i, s_j) =0 \ \forall\, i \neq j$ *and cov* $(w_i, s_j) = 0 \ \forall\, i,j$.
so :

$$\begin{cases} f_0(y) = P(y/H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_s{}^2}} e^{-\frac{1}{2}\frac{y_i{}^2}{\sigma_s{}^2}} \\ f_1(y) = P(y/H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\sigma_s{}^2+\sigma_w{}^2)}} e^{-\frac{1}{2}\frac{y_i{}^2}{\sigma_s{}^2+\sigma_w{}^2}} \end{cases} \qquad (5)$$

$$P_F = P_r\left(\sum_{i=1}^{N} y_i{}^2 > \gamma\backslash H_0\right) = P_r\left\{\sum_{i=1}^{N}\left(\frac{y_i}{\sigma_w}\right)^2 > \frac{\gamma}{\sigma_w^2}/H_0\right\} \qquad (6)$$

where $\sum_{i=1}^{N}\left(\frac{y_i}{\sigma_w}\right)^2$ follows a noncentral chi-squared distribution with N degrees of freedom. Let us make: $X = \sum_{i=1}^{N}\left(\frac{y_i}{\sigma_w}\right)^2$ then the distribution probability is given by:

$$P(X/H_0) = \frac{1}{2^{\frac{N}{2}}.\Gamma\left(\frac{N}{2}\right)} X^{\frac{N}{2}-1} e^{-\frac{X}{2}} \qquad (7)$$

where $\Gamma(a) = \int_0^{+\infty} t^{a-1}e^{-t}\,dt$ is the gamma function.

Similarly for P_D:

$$P_D = P_r\{\sum_{i=1}^{N} {y_i}^2 > \gamma \,|H_1\} = P_r\left\{\sum_{i=1}^{N}\left(\frac{y_i}{\sqrt{\sigma_w^2+\sigma_s^2}}\right)^2 > \frac{\gamma}{\sigma_w^2+\sigma_s^2}/H_1\right\} \qquad (8)$$

where $\sum_{i=1}^{N}\left(\frac{y_i}{\sqrt{\sigma_w^2+\sigma_s^2}}\right)^2$ follows a noncentral chi-squared distribution with N degrees of freedom. Also, the distribution probability is given by:

$$P_D = 1 - \Gamma_{inc}\left(\frac{\gamma}{2(\sigma_w^2+\sigma_s^2)}, \frac{N}{2}\right) \qquad (9)$$

We will define the incomplete gamma used by Matlab function:

$$\Gamma_{inc}(x,a) = gammainc(x,a) = \frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}dt$$

This allowed derive, analytically, the minimum number of samples that are required to complete a prescribed performance (P_F, P_d) for a given SNR, the expression of N is given as follows [18]:

$$N \approx \lceil [SNR^{-1}\, Q^{-1}(P_F) - (1 + SNR^{-1})Q^{-1}(P_D)]^2 \rceil \qquad (10)$$

for sufficiently large number N [17].

3.3. Power Spectral Estimation

The energy detection is the common used tool for spectrum sensing due its low computational cost and implementation complexity. It is designed to decide the presence or the absence of UP without a priori knowledge of statistical characteristics of the primary signal [19].

There exist two domains for implementing the energy detection technique, in time as well as frequency domain by the use of the Fast Fourier transform (FFT).

If the Fourier transform X (f) of x(t) signal exists then according to the theorem of Perceval:

$$\int_{-\infty}^{+\infty}|x(t)|^2 dt = \int_{-\infty}^{+\infty}|X(f)|^2\, df \qquad (11)$$

Signal energy is conserved in both time and frequency domains however the representation in the frequency domain is more flexible [20].

3.3.1. The Periodogram:

It is the simplest non-parametric method of power spectral estimation. The spectral power density (PSD) of length L of the signal $x_L(n)$ is defined by [21-22]:

$$P_{xx}(f)=\frac{1}{L}\left|\sum_{n=0}^{L-1} x_L(n)e^{-j2\pi fn}\right|^2 \quad (12)$$

3.3.2. The average Periodogram (Welch):

The main problem with the Periodogram is high variance (inconsistency). A simple solution for this is to apply the average of a set of estimates (assumed to be independent). The signal of length N is divided into k segments overlapped with length L, and then each segment periodogram is calculated; the average Periodogram I estimated as follows [8]:

$$P_{AVER}(f)=\frac{1}{K}\sum_{m=0}^{K-1} P_{PER}(f)^{(m)} \quad (13)$$

Where

$$P_{PER}(f)^{(m)}=\frac{1}{L}\left|\sum_{n=0}^{L-1} x_L(n)e^{-j2\pi fn}\right|^2 \quad (14)$$

4. THE ENERGY DETECTION ALGORITHM

4.1. The Implementation of Energy Detection Algorithm

It includes 5 main phases:

1) Initialization: entering the values of P_D and P_F, the duration and the number of signals and the sampling frequency.

2) The primary signal generation: entering the F_i frequency and make the summation of primary signals by adding white Gaussian noise of diverse SNR values;

3) Calculation of the threshold according to the equations (2) and (10);

4) Estimation of the PSD: by using the 2 techniques of spectral estimation of power (Periodogram and Welch);

5) Evaluation: Comparison with the threshold calculated in step 3) and decision making.

4.2. Simulation and Result

We generate at maximum 6 sinusoidal signals with frequencies 1 kHz, 2 kHz, 3 kHz, 4 kHz, 5 kHz, and 6 KHz. The sampling frequency is F_s = 14KHz. In this work we have taken the respective values of P_D and P_F equal to 0.95 and 0.05. The following table 1 shows the results obtained for different values of SNR and durations of observation.

The Periodogram method can be a useful tool for spectral estimation in case of high SNR more specifically where there data is longer; where it arises the importance of the observation period. Table 2 below shows the results of the study of the detection performance depending on the duration of observation (Figure 4).

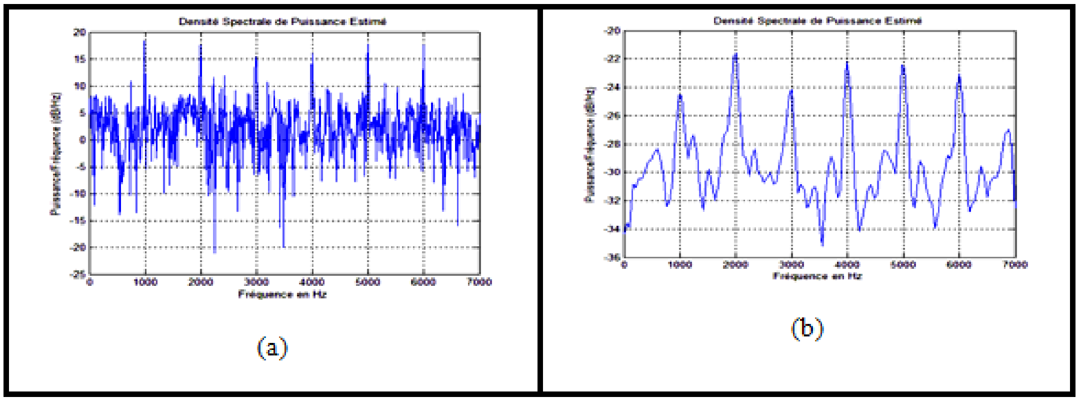


Figure 4. Power spectral density PSD estimation for an SNR of -4 dB using:
(a) Periodogram; (b) the Welch method

Table 1. Estimation of the DSP for Welch VS. Periodogram.

SNR (dB)	Duration (ms)	Estimation of PSD	True Detection	False Detection
-4	0.0651	Welch	4/6	2/6
		Periodogram	5/6	1/6
	0.1	Welch	5/6	1/6
		Periodogram	5/6	1/6
4	0.0651	Welch	5/6	1/6
		Periodogram	4/6	2/6
	0.1	Welch	5/6	1/6
		Periodogram	6/6	0/6

Table 2. The performance of detection on the basis of the observation period.

SNR (dB)	Duration	True Detection	False Detection
-4	0.9 ms	4/6	2/6
	1.6 ms	6/6	0
0	0.1 ms	6/6	0
	0.6 ms	6/6	0
4	0.1 ms	6/6	0
	0.6 ms	5/6	1/6

The previous synthesis shows that the detection performance depend not only on the probability of false alarm and detection, but it also depends on the optimal duration of observation T_{opt}. We note that the time required for a low SNR is wider than that for a high SNR. Thus, we can confirm that the required observation period is inversely proportional to the level of the SNR.

5. COGNITIVE RADIO BASED ON THE DETECTION OF ENERGY

After the implementation of the algorithm of the identification of the free spectral bands, for a given scenario, based energy detection technique; we proceed, in the present phase, to the context of cognitive radio. In other words, we simulate a transmission chain of an alphanumeric message via a noisy Gaussian channel by identifying free spectral bands to be allocated by secondary users (cognitive radio).

5.1. Description of the Algorithm

In our example, we generate an alphanumeric message converted to 8 bit ASCII encoding. Next initialization phase comprises the duration of the signal T, carrier frequencies, the value of P_D and P_F and the secondary signal. The used modulation is BPSK (Binary Phase Shift Keing).

The modulated primary signal is transmitted through an AWGN channel. Primary signals are determined by the technique of the ED and as the detector output is above the threshold then it says that the UP is present and vice versa. The insertion of secondary user is introduced in the band where the primary user is absent. Figure 5 illustrates the methodology for implementing the detection technique in the context of cognitive radio.

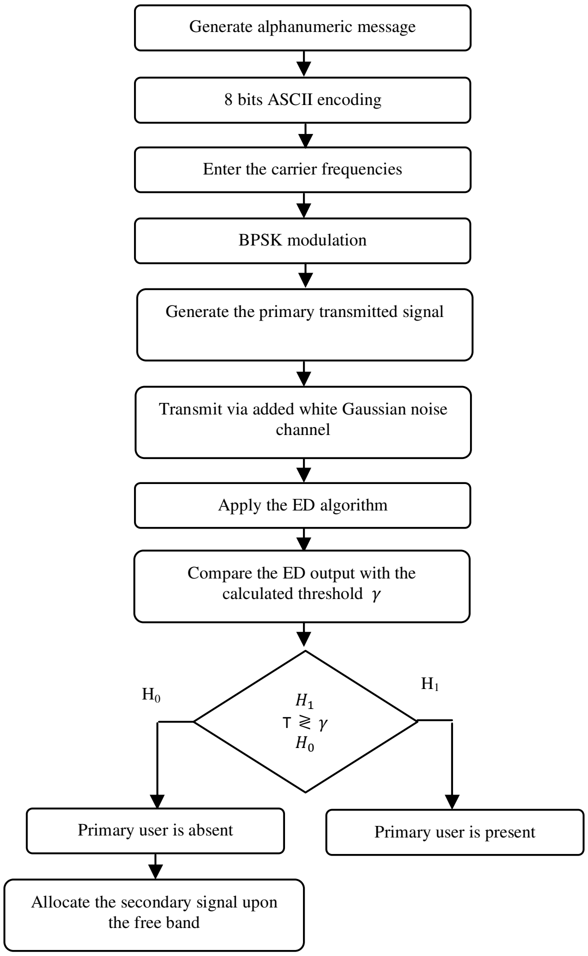

Generate alphanumeric message
8 bits ASCII encoding
Enter the carrier frequencies
BPSK modulation
Generate the primary transmitted signal
Transmit via added white Gaussian noise channel
Apply the ED algorithm
Compare the ED output with the calculated threshold γ
H_0
H_1
H_1
$\mathrm{T} \gtrless \gamma$
H_0
Primary user is absent
Primary user is present
Allocate the secondary signal upon the free band

Figure 5. Methodology for implementing ED in the context of RC.

5.2. Results and discussions

In our example, we generate a randomly alphanumeric message of 19 characters; this message will be ASCII-encoded 8-bit. Considering 3 BPSK modulated primary signals with carrier frequencies of 2.5 KHz, 6.5 KHz and 10.5 KHz. The sampling frequency is F_s = 24 kHz and the duration of signal is t = 0.08ms. Assuming that there are two primary users 2.5 KHz and 10.5KHz frequencies are present and the primary user with a frequency of 6.5KHz is absent. The signal is transmitted through an AWGN channel.

Calculation of the threshold: The threshold is calculated based on equation (2). The number of observed samples N is calculated from (5). We choose P_D = 0.95, P_F = 0.05, SNR = - 2DB and duration of signal is t = 0.08ms. The cognitive radio system looks permanently for the hole of the spectrum where the primary user is absent which is determined by the method of energy detection. When it finds the hole of the spectrum, immediately it attributes it to the secondary user (US). Figure 6 shows the occupation of the unused bandwidth by secondary user.

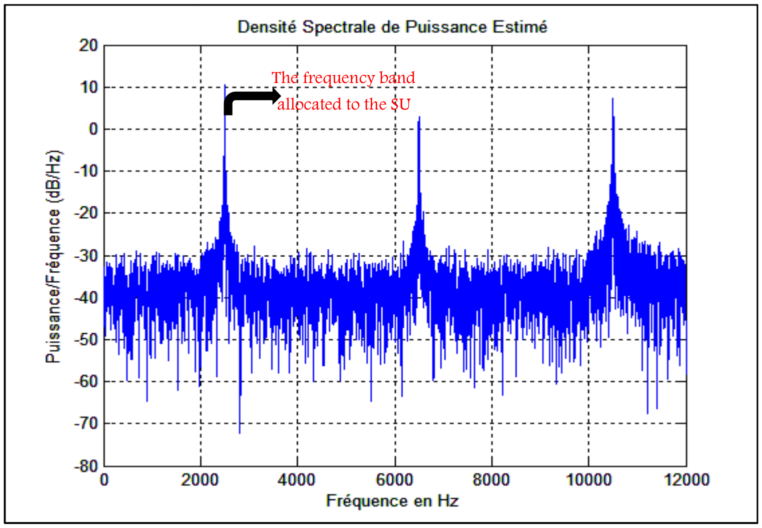


Figure. 6 Allocation of secondary user in the free band

Figure 7 shows the estimation of the spectral power density by the technique of Periodogram for an SNR = - 30dB. Unfortunately, for this very low level SNR, the algorithm of the ED fails to identify the free spectral bands due the high number of representative noise within the informative signal peaks.

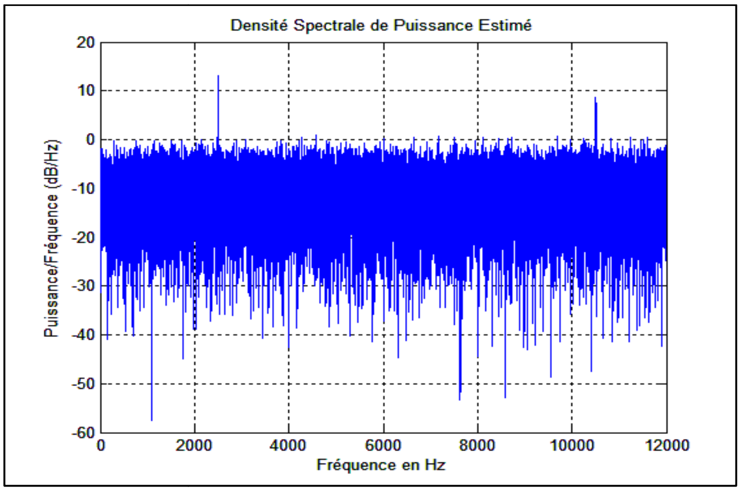


Figure 7. The PSD estimated by the technique of periodogram for an SNR =-30dB

6. CONCLUSIONS

In this paper, the technique of detection of energy has been introduced in the context of cognitive radio. The realized algorithm is based on the Neyman-Pearson test and calculation of P_F and P_F. we have implemented an energy detection algorithm. In our example it was shown that the detection of energy performance also depends on the optimal observation duration, and the method chosen for the estimation of the power spectral density (PSD). An example of telecommunications systems that provides a scenario for the RC system was exposed, and also was given a developed procedure that allows to insert the secondary user in the unoccupied band of opportunistically in a dynamic way allowing better allocation of available frequency resources. The signal to emit, is modulated in BPSK is transmitted through an AWGN channel. The presence / absence of primary user is determined by the technique of energy detection (ED), when the primary user is absent, the band is assigned to the secondary user.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) “This is my paper”, ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

[1] J. Mitola, A. Attar, H. Zhang, O. Holland, H. Harada, and H. Aghvami, “Special issue on achievements and the road ahead: the first decade of cognitive radio,” IEEE Trans. Veh. Technol., vol. 59, no. 4, May 2010.

[2] J. Ma, G. Y. Li, and B. H. Juang, “Signal processing in cognitive radio,” Proc. IEEE, vol. 97, no. 5, pp. 805–823, May 2010.

[3] S. Haykin, D. J. Thomson, and J. H. Reed, “Spectrum sensing for cognitive radio,” Proc. IEEE, vol. 97, no. 5, pp. 849–877, May 2010.

[4] S. H. Song, K. Hamdi, and K. B. Letaief, “Spectrum sensing with active cognitive systems,” IEEE Trans. Wireless Commun., vol. 9, pp. 1849–1854, June 2010.

[5] Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, and Shantidev MohantyA, « Survey on Spectrum Management in Cognitive Radio Networks », Georgia Institute of Technology, IEEE Communications Magazine, April 2008.

[6] Tevfik Yucek and Huseyin Arslan, «A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications », IEEE Communications Surveys & Tutorials, Vol. 11, No. 1, First Quarter 2009.

[7] Christophe MOY "Evolution de la conception radio : de la radio logicielle à la radio intelligente" Habilitation à Diriger des Recherches (HDR), Université de Rennes 1, 8 octobre 2008

[8] B. Wang and K. J. R. Liu, “Advances in cognitive radio networks: a survey,” IEEE J. Sel. Topics Signal Process., vol. 5, no. 1, pp. 5–23, Feb 2011.

[9] Mansi Subhedar et Gajanan B, «Spectrum Sensing Techniques in cognitive radio network: A survey », International Journal of Next Generation Network Vol.3, No.2, 37-51 June 2011.

[10] Phunchongharn, P.; Hossain, E.; Niyato, D.; Camorlinga, S. A cognitive radio system for e-health applications in a hospital environment Wireless Communications, IEEE February 2010 Volume: 17 Issue:1 On page(s): 20 - 28 ISSN: 1536-1284

[11] Chavez-Santiago, R.; Balasingham, I.; 2011 IEEE 16th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 10-11 June 2011 ; 148 – 152 Kyoto E-ISBN: 978-1-61284-280-6 Print ISBN: 978-1-61284-281-3

[12] Joseph Mitola III, « Cognitive Radio an Integrated Agent Architecture for Software Defined Radio », Royal Institute of Technology (KTH), May 2000.

[13] Architecture for Cognitive Radio Networks Broadband Wireless Networking Lab School of Electrical and Computer Engineering Georgia Institute of Technology http://www.ece.gatech.edu/research/labs/bwn/CR/projectdescription.html

[14] I. F. Akyildiz, W. Y. Lee, M.C. Vuran and S. Mohanty, "NeXt Generation / Dynamic Spectrum Access / Cognitive Radio Wireless Networks: A Survey,"Computer Networks Journal (Elsevier), Vol. 50, pp. 2127-2159, September 2006

[15] Gyanendra Prasad Joshi , Seung Yeob Nam and Sung Won Kim Review Cognitive Radio Wireless Sensor Networks: Applications, Challenges and Research Trends Sensors 2013, 13(9), 11196-11228; doi:10.3390/s130911196 http://www.mdpi.com/1424-8220/13/9/11196/htm

[16] Mahmood A. Abdulsattar and Zahir A. Hussein, «Energy detection technique for Spectrum sensing in cognitive radio: a survey», International Journal of Computer Networks & Communications (IJCNC) Vol.4, No.5, September 2012.

[17] Ratsirarson s. a, Rakotonirina t, Ravonimanantsoa n.m.v, «Analyse Des Différents Types Des Fonctions De La Radio Cognitive », Université d’Antananarivo BP 1500, Ankatso – Antananarivo 101 – Madagascar.

[18] H. Urkowitz, «Energy Detection of Unknown Deterministic Signals», Proc. of the IEEE, vol. 55, no. 4, pp. 523-531, 1967.

[19] A.M. Fanan, N.G. Riley, M. Mehdawi, M. Ammar, and M. Zolfaghari, «Survey: A Comparison of Spectrum Sensing Techniques in Cognitive Radio», Int'l Conference Image Processing, Computers and Industrial Engineering (ICICIE'2014) Jan. 15-16, 2014 Kuala Lumpur (Malaysia).

[20] Ahmad Ali Tabassam et Muhammad Uzair Suleman, «Building Cognitive Radios in MATLAB Simulink – A Step Towards Future Wireless Technology », Wireless Advanced, IEEE,2011.

[21] J.McNames, «Spectral Estimation », Portland State University ECE 538/638Ver. 1.15.

[22] Shan He et Hongmei Gou, «Spectral Estimation», ENEE 624 Advanced Digital Signal Processing, University of Maryland at College Park, Dec. 16, 2002.

AUTHORS

Professor CHOUAKRI, SID AHMED, engineering educator, researcher; married. PhD (hon.), U. Tlemcen, Algeria, 2008. Rschr., Prof. U. Sidi Bel Abbès, Algeria, since 1997. Member of Laboratory of Telecommunications and Digital Signal Processing at the University of Sidi Bel Abbès, Algeria. He received the engineering degree in Computer Science from the National Institute of electricity and electronics of Boumerdes, Algeria in 1992 and the Master degree and PhD in Electronics from the University of Tlemcen, Algeria in respectively, 1997 and 2008.

Domains of interest: biomedical engineering, signal processing (digital filtering, time-frequency analysis, wavelet transform…), Hardware systems, real time systems, wireless for telemedicine, telecardiology, RFID systems, Cognitive radio…

AUTHOR INDEX