Computer Science & Information Technology 40

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan (Eds)

Computer Science & Information Technology

International Conference on Computer Science and Information Technology (CSTY 2015)
Bangalore, India, April 25 ~ 26 - 2015

AIRCC

Volume Editors

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

Sundarapandian Vaidyanathan,
R & D Centre,
Vel Tech University, India
E-mail: sundarvtu@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-37-3
DOI : 10.5121/csit.2015.50901 - 10.5121/csit.2015.50907

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

Preface

The International Conference on Computer Science and Information Technology (CSTY 2015) was held in Bangalore, India, during April 25~26, 2015. The International Conference on Signal and Image Processing (SIGI 2015), The International Conference on Artificial Intelligence and Applications (AI 2015), and The International Conference of Managing Value and Supply Chains (MaVaS 2015) were collocated with the CSTY-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 CSTY-2015, SIGI-2015, AI-2015, MaVaS-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, CSTY-2015, SIGI-2015, AI-2015, MaVaS-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 CSTY-2015, SIGI-2015, AI-2015, MaVaS-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.

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan

Organization

General Chair

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

Program Committee Members

Abdolreza Hatamlou	Islamic Azad University, Iran
Abe Zeid	Northeastern University, USA
Abraham Sanchez Lopez	Autonomous University of Puebla, Mexico
Ahmed Y. Nada Aqu	Al-Quds University, Palestine
Alaa Hussein Al-hamami	Amman Arab University, Jordan
Ali Abid D. Al-Zuky	University of Baghdad, Iraq
Amol D Mali	University of Wisconsin-Milwaukee,USA
Anil Kumar Dubey	Govt. Engineering College Ajmer
Ankit Chaudhary	Truman State University, USA
Arifa Ferdousi	Varendra University, Bangladesh
Ashraf A. Shahin	Cairo University, Egypt
Ayad Ghany Ismaeel	Erbil Polytechnic University, Iraq
Barbaros Preveze	Cankaya University, Turkey
ChanChristine	University of regina, Canada
Chin-Chih Chang	Chung Hua University, Taiwan
Dac-Nhuong Le	Hai Phong University, Vietnam
Derya Birant	Dokuz Eylul University, Turkey
El-Sayed M. EL-Rabaie	Menouf, Egypt
Farshchi	Tehran University, Iran
Fatih Korkmaz	Cankiri karatekin university, Turkey
Grienggrai Rajchakit	Maejo University, Thailand
Hamid Mcheick	Universite du Quebec a Chicoutimi,Canada
Hanan Salam	University of Pierre and Marie Curie, France
Hassini Noureddine	University of Oran , Algeria
Hossein Jadidoleslamy	MUT University, Iran
Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University, Egypt
Israa SH.Tawfic	Gaziantep University, Turkey
Israa Shaker Tawfic	Ministry If Scienece And Technology, Iraq
Israashaker Alani	Ministry of Science and Technology, Iraq
Kai-Long Hsiao	Taiwan Shoufu University, Taiwan
Kavitha Rajamani	St. Aloysius College, India
Kenneth MAPOKA	Iowa state university, USA
Krishna Prakash K	M.I.T. Manipal, India
M.Chithirai Pon Selvan	Manipal University,Dubai
Mahdi Mazinani	Azad University, Iran
Manal	King Abdulaziz University, South Africa

Manish Mishra	Haramaya University, Ethiopia
Masoud ziabari	Mehr Aeen University, Iran
Metin Soycan	Yildiz Technical University,Turkey
Mohammad Farhan Khan	University of Kent, United Kingdom
Mohammad Masdari	Islamic Azad University, IRAN
Mohammed Erritali	Sultan Moulay Slimane University,Morocco
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Mujiono Sadikin	Universitas Mercu Buana, Indonesia
Naga Raju	Acharya Nagarjuna University, India
Natarajan Meghanathan	Jackson State University, USA
Nouriii Nouri.Naz	Innov'com Lab Sup'com, Tunisie
Ognjen Kuljaca	Brodarski Institute, Croatia
Oussama Ghorbel	University of Troyes, Tunisia
Peiman Mohammadi	Islamic Azad University, Iran,
Rahil Hosseini	Islamic Azad University, Iran
Ramayah T	Universiti Sains Malaysia, Malaysia
Reda Mohamed Hamou	Tahar Moulay University of Saida, Algeria
Rekha Mehra	Govt. Engineering College Ajmer
Reza Ebrahimi Atani	University of Guilan, Iran
RHATTOY Abdallah	Moulay Ismail University, Morocco
Rim Hadded	Sup'com, Tunisia
Saad M. Darwish	Alexandria University, Egypt
Saeed Tavakoli	University of Sistan and Baluchestan, Iran
Salem Hasnaoui	Al-Manar University, Tunisia
Samadhiya	National Chiao Tung University, Taiwan
Sanjay K. Dwivedi	Babasaheb Bhimrao Ambedkar University, India
Sarah M. North	Kennesaw State University, USA
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Shaik Basheera	Eswar college of engineering, India
Shengxiang Yang	De Montfort University, UK
Stefano Berretti	University of Florence, Italy
Upendra Kumar	Mahatma Gandhi Institute of Technology, India
Viliam Malcher	Comenius University, Europe
Zebbiche Toufik	University of Blida 1,Algeria

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

International Conference on Computer Science and Information Technology (CSTY 2015)

International Conference on Signal and Image Processing (SIGI 2015)

International Conference on Artificial Intelligence and Applications (AI 2015)

International Conference of Managing Value and Supply Chains (MaVaS 2015)

ON THE DISTRIBUTION OF THE MAXIMAL CLIQUE SIZE FOR THE VERTICES IN REAL-WORLD NETWORK GRAPHS AND CORRELATION STUDIES

Natarajan Meghanathan

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

ABSTRACT

The high-level contributions of this paper are as follows: We modify an existing branch-and-bound based exact algorithm (for maximum clique size of an entire graph) to determine the maximal clique size that the individual vertices in the graph are part of. We then run this algorithm on six real-world network graphs (ranging from random networks to scale-free networks) and analyze the distribution of the maximal clique size of the vertices in these graphs. We observe five of the six real-world network graphs to exhibit a Poisson-style distribution for the maximal clique size of the vertices. We analyze the correlation between the maximal clique size and the clustering coefficient of the vertices, and find these two metrics to be poorly correlated for the real-world network graphs. Finally, we analyze the Assortativity index of the vertices of the real-world network graphs and observe the graphs to exhibit positive assortativity with respect to maximal clique size and negative assortativity with respect to node degree; nevertheless, we observe the Assortativity index of the real-world network graphs with respect to both the maximal clique size and node degree to increase with decrease in the spectral radius ratio for node degree, indicating a positive correlation between the maximal clique size and node degree.

KEYWORDS

Maximal Clique Size, Node Degree, Correlation, Assortativity Index, Distribution, Network Graphs, Clustering Coefficient

1. INTRODUCTION

Network Science is an emerging area of research interest to study complex real-world networks from a graph theoretic point of view. We abstract the complex network as a graph with the nodes representing the vertices and the connections between any two nodes in the network modeled as edges in the graph. It is imperative that the algorithms run on these large scale graphs be as efficient as possible and do not take significant time to determine the metrics of interest. Though there exists efficient polynomial-time algorithms to determine widely studied metrics [1] like centrality, diameter, clustering coefficient, etc on these graphs, there still exists certain metrics like clique such that the problem of determining a maximum size clique is NP-hard [2]. A clique on a graph is a subset of the vertices such that there exists an edge between any two vertices in this subset; an algorithm to find cliques of various sizes could be used to identify closely-knit communities [3-5] of various sizes (constituent nodes) in real-world network graphs.

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 01–13, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50901

The "maximum size clique" for a graph of n vertices is a clique of the largest size k ($k \leq n$) such that there does not exist a clique of size $k + 1$ in the graph. A "maximal size clique for a vertex i" in a graph is the clique of the largest size that involves vertex i as one of the constituent vertices. While the maximum size clique for a graph is the maximal size clique for its constituent vertices, there could exist several other vertices in the graph for which the maximal size clique is smaller than the maximum size clique. Most of the research focus in the literature is to develop exact algorithms that could determine the maximum size clique for the entire graph as efficiently as possible with respect to both time and space complexity. Very little attention has been made to determine the maximal size cliques for the individual vertices in the graph. Specifically, to the best of our knowledge, no attempt has been made to analyze the distribution of the maximal clique sizes of the individual vertices in the real-world complex network graphs. In this paper, we choose a recently proposed exact algorithm [6] to determine the size of the maximum clique for large-scale complex network graphs and extend it to determine the size of the maximal clique that a particular node is part of. Using the exact algorithm to determine maximal clique size for the individual vertices of the graphs, we determine the distribution of the maximal clique size for all the six real-world network graphs considered in this study. We observe that five of the six real-world network graphs (irrespective of their number of nodes and degree distribution) exhibit a Poisson-style distribution for the maximal clique size; this is a significant observation that has not been hitherto reported in the literature.

The second half of our paper focuses on identifying a computationally-light metric for the individual nodes of a graph that correlates well (either positively or negatively) to that of the maximal clique size (which we categorize as a computationally-hard metric, owing to the NP-hard nature of the problem to determine this metric and the significant time complexity involved in the exact algorithms for this metric). Once we identify such a computationally-light metric that correlates well with the maximal clique size of the vertices in complex network graphs, we could infer a ranking of the vertices based on this computationally-light metric as a ranking of the vertices based on the maximal clique size. To the best of our knowledge, we have not come across any such study to identify a computationally-light metric that correlates well with the maximal clique size for real-world network graphs. Ours is the first attempt in this direction. The two candidate computationally-light metrics that we consider are the clustering coefficient and the node degree. The clustering coefficient of a vertex is the ratio of the number of edges between the neighbors of the vertex to that of the maximum number of edges possible between the neighbors of the vertex. Our conjecture is that nodes that are part of a larger clique are more likely to have a larger clustering coefficient and vice-versa. Similarly, we conjecture that nodes that have a larger degree (number of neighbors) are likely to be part of cliques of larger size and vice-versa. Results of our correlation studies on real-world network graphs reveal that the maximal clique size has good correlation with node degree (especially as the variation in the node degree increases), whereas the maximal clique size correlates poorly with the clustering coefficient. We further confirm the positive correlation between the maximal clique size and node degree through an analysis of the Assortativity index of the vertices [1] in the real-world network graphs with respect to these two metrics. We observe the real-world network graphs could be ranked in a similar order in the decreasing order of the Assortativity index of the vertices with respect to both the maximal clique size and the node degree.

The rest of the paper is organized as follows: Section 2 describes related work on analysis of complex network graphs using cliques. Section 3 describes an efficient exact algorithm to determine the maximum clique size for an entire graph and our extension to determine the maximal clique size for the individual vertices of the graph. Section 4 presents the real-world network graphs studied in this paper and an analysis of their degree distribution and distribution of the maximal clique size of the vertices. Section 5 presents the results of the correlation studies between the maximal clique size and clustering coefficient. Section 6 presents the results of the correlation studies between the maximal clique size and the node degree. Section 7 presents the results of Assortativity index-based analysis of the real-world network graphs with respect to maximal clique size and node degree. Section 8 concludes the paper. Throughout the paper, we use the terms 'node' and 'vertex', 'link' and 'edge' interchangeably. They mean the same.

2. RELATED WORK

The research focus with regards to cliques in the context of complex networks is to come up with efficient heuristics to reduce the run-time complexity in determining the maximum size clique for the entire network graph. Though branch-and-bound has been the common theme among these works, the variation is in the approach used to arrive at the bounds and enforce them in the search space. Strategies used for pruning the search space are typically based on node degree (e.g., [6]), vertex ordering (e.g., [7]) and vertex coloring (e.g., [8]). Recently, a parallelized approach [9] for branch and bound has also been proposed for determining cliques in real-world networks ranging from 1000 to 100 million nodes. Nevertheless, none of the research so far has focused on identifying correlation between the maximal clique size for an individual vertex (the size of the largest clique that a particular vertex is part of) with any of the commonly studied metrics (like node degree, clustering coefficient) for network analysis. Ours is the first step in this direction. With the problem of determining maximum size clique for the entire network graph and maximal size cliques for the individual vertices being NP-hard and computationally time-consuming for complex real-world networks of larger size, it becomes imperative to analyze the correlations of the maximal clique size values of the individual vertices with that of the network metrics that can be easily computed so that meaningful inferences about maximal clique size values can be made.

3. CLIQUE

A clique is a sub graph of a graph in which all the vertices are adjacent to each other. The problems of finding maximum size clique for the entire graph as well as the maximal size cliques for the individual nodes are NP-hard problems [2]. Several exact algorithms (that at the worst case incur exponential time for a NP-hard problem) have been proposed to determine maximum size cliques for sparse graphs. Recently, with the surge in interest to analyze large real-world networks from a graph theoretic point of view, researchers have proposed efficient exact algorithms (e.g., [6-9]) to determine maximum size cliques for large/dense graphs. The common theme [10] behind these algorithms is a branch and bound approach of searching through all possible candidate cliques and limiting the search to only viable candidate sets of vertices whose agglomeration has scope of being a clique of size larger than the currently known clique found as part of the search; the variation among these exact algorithms is the pruning strategy (the approach taken to compute the bounds and use them) to limit the search. In this section, we will describe one such branch and bound-technique based exact algorithm that has been recently proposed in the literature [6] to determine maximum size clique in large network graphs and explain our modification to the algorithm so that it can be used to determine the maximal cliques that each vertex in the graph is part of; the largest among these cliques is the maximum size clique for the entire graph.

Figure 1 outlines the pseudo code of the algorithm (proposed originally in [6]) to determine the maximum size clique for an entire graph. The algorithm starts with an estimate of 0 for the maximum size clique (variable *max*) in the entire graph; the value for *max* is updated as and when a clique of size larger than the latest value of *max* is found. The procedure MAXCLIQUE proceeds in iterations, with each iteration designed to determine the maximum size clique for the entire graph that could also include vertex v_i (considered in the increasing order of the IDs). In a particular iteration, vertex v_i is considered worthy of exploration for presence in a maximum size clique only if its degree is at least the value of *max* at that time (i.e., only vertices that could be part of a clique of size larger than the currently known maximum size clique are considered - a pruning strategy). For each such vertex v_i, a candidate set U of neighbor vertices v_j (whose degree is at least the latest value for *max*) is constructed and passed to the sub routine CLIQUE to find a clique among these vertices; the initial size of the clique is 1 - accounting for v_i.

```
Procedure MAXCLIQUE ($G = (V, E)$)
    $max \leftarrow 0$
    for $i$ : 1 to $|V|$ do
        if degree($v_i$) $\geq max$ then
            $U \leftarrow \phi$
            for each $v_j \in$ Neighbor($v_i$) do
                if degree($v_j$) $\geq max$ then
                    $U \leftarrow U \cup \{v_j\}$
            CLIQUE($G$, $U$, 1)
```

```
Subroutine CLIQUE($G = (V, E)$, $U$, $size$)
  // $size$ is the size of clique found so far
  if $U = \phi$ then
    if $size > max$ then
      $max \leftarrow size$
    return
  while $|U| > 0$ do
    if $size + |U| \leq max$ then
      return
    select any vertex $u$ from $U$
      $U \leftarrow U \backslash \{u\}$
      $N'(u) := \{w \mid w \in$ Neighbor($u$) $\wedge$
                    degree($u$) $\geq max\}$
      Clique($G$, $U \cap N'(u)$, $size + 1$)
```

Figure 1. Exact Algorithm to Determine Maximum Size Clique for a Graph (adapted from [6])

The sub routine CLIQUE called with vertex v_i as the first constituent vertex of the largest possible clique involving v_i, expands with one vertex at a time through a combination of iterations and recursions; the sub routine runs as long as the size of the set U is greater than zero or if the current value of *max* is less than the sum of the sizes of the set U and the current clique found so far (a pruning strategy). In each such iteration, a vertex u (that is also a neighbor of the starting vertex v_i and the other vertices in the clique determined so far) is randomly removed from the set U and the neighbors of u that are also present in U (and hence are neighbors of the starting vertex v_i and the other vertices that are part of the clique found so far) are only further considered to be candidates that could be part of the clique, and a recursive call to the CLIQUE sub routine is made with the value of variable *size* (the size of the largest clique found so far involving vertex v_i) incremented by 1 - accounting for u as the latest entrant in the clique determined so far. Each recursive call to CLIQUE is accompanied by an iteration where a vertex u (that is also a neighbor of the vertices already part of the clique) is removed from the set U passed to the sub routine and only the neighbors of u that are also neighbors of the vertices already in the clique are considered. During any such recursive call, if the size of the set U passed to the sub routine CLIQUE reaches zero, the algorithm terminates the sequence of recursions and updates the value of *max* if the size of the clique determined so far involving vertex v_i is larger than the current value of *max*. During the sequence of returns from the recursive calls, it is possible that a new sequence of recursions and iterations is triggered due to the presence of a neighbor u of v_i that has scope for being in a clique (involving v_i) of size larger than the clique found so far for the entire graph. The algorithm explores all such possible cliques involving vertex v_i that have scope for exceeding the currently known maximum size clique for the entire graph.

At the end, the algorithm returns the maximum size clique for the entire graph that also happens to be the maximal size clique involving some vertex v_i such that there is no other vertex v_j ($i > j$) that is also part of the clique. Since the algorithm proceeds with vertices in the increasing order of their IDs, if the maximum size clique for the entire graph involves at least one vertex v_i with a smaller ID, the presence of the maximum size clique is detected much earlier and the subsequent iterations (with vertices whose IDs are greater than v_i, but could be part of only cliques of size smaller or equal to the maximum size clique of the entire graph involving v_i) are merely pruned, contributing to the time-efficiency of the algorithm. Hence, the labeling of the vertices with their IDs plays a significant role in the run-time complexity of the algorithm; the algorithm is capable of quickly determining the maximum size clique if the latter comprises of at least one vertex with a smaller ID.

Figure 2 illustrates our modifications (to determine the size of the maximal clique that each vertex is part of) to the pseudo code of the algorithm presented in Figure 1. The tradeoff is an increase in the run-time of the algorithm: we cannot just prune our search based on the vertex IDs; we have to explore the neighborhood of each of the vertices to determine the maximal size clique that each vertex is part of. Since to start with, the maximal size clique known for vertex v_i is 0, there is no need to filter the neighbors of v_i in procedure MAXIMALCLIQUE based on the degree of the neighbors; all neighbors of v_i are included in the set U and passed onto the sub routine CLIQUE. However, we could retain all of the pruning steps in sub routine CLIQUE (called to find the maximal size clique for each of the vertices v_i) and recursive calls to the same: there is no need to explore the neighbors of vertex u whose degree is less than that of the currently known maximal clique size for vertex v_i.

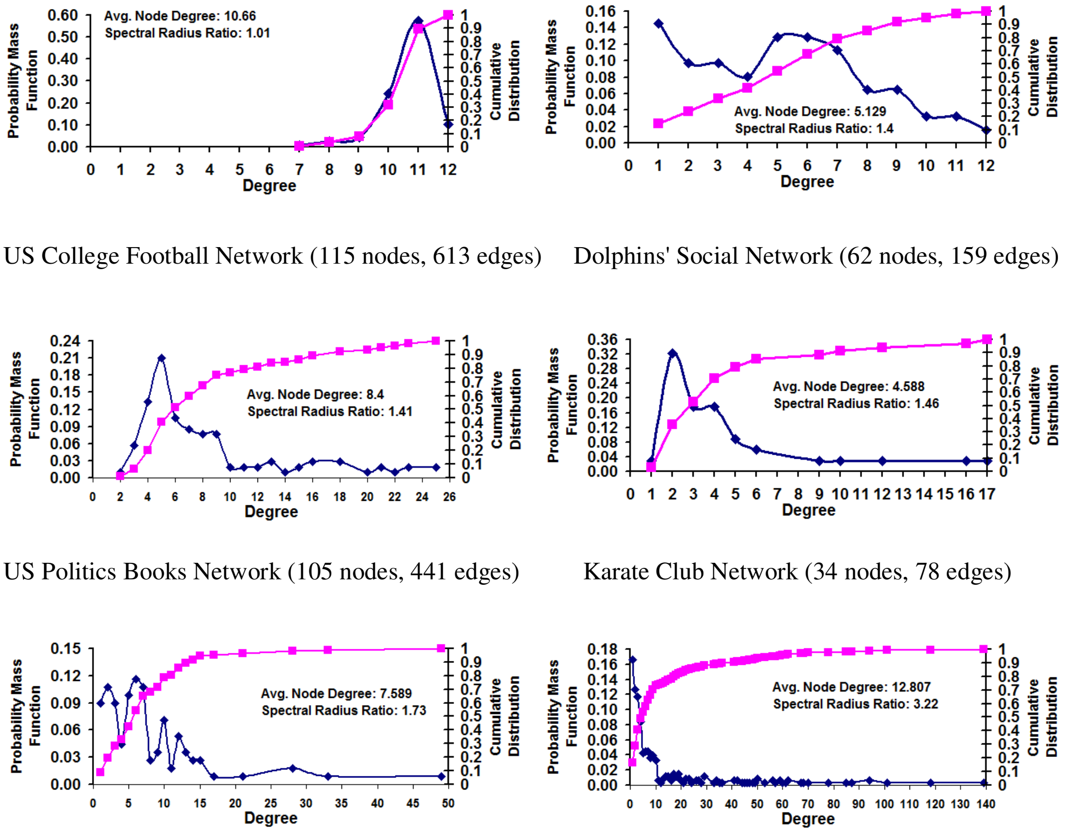
```
Procedure MAXIMALCLIQUE ($G = (V, E)$)
    for $i$ : 1 to $|V|$ do
        $maximalCliqueSize[v_i] \leftarrow 0$
        $U \leftarrow \phi$
        for each $v_j \in$ Neighbor($v_i$) do
            $U \leftarrow U \cup \{v_j\}$
        CLIQUE($G$, $v_i$, $U$, 1)

Subroutine CLIQUE($G = (V, E)$, $v_i$, $U$, $size$)   // $size$ is the size of clique found so far for vertex $v_i$
  if $U = \phi$ then
    if $size > maximalCliqueSize[v_i]$ then
        $maximalCliqueSize[v_i] \leftarrow size$
    return
  while $|U| > 0$ do
     if $size + |U| \leq maximalCliqueSize[v_i]$ then
         return
     select any vertex $u$ from $U$
     $U \leftarrow U \setminus \{u\}$
     $N'(u) := \{w \mid w \in$ Neighbor($u$) $\wedge$ degree($u$) $\geq maximalCliqueSize[v_i]\}$
     Clique($G$, $v_i$, $U \cap N'(u)$, $size + 1$)
```

Figure 2.Exact Algorithm to Determine the Maximal Clique Size for each Vertex in a Graph (adapted from [6])

4. Real-World Network Graphs And Their Analysis

In this section, we describe the network graphs analyzed and illustrate the degree distribution and the distribution of the maximal clique size of the vertices in the network graphs. We do so to understand the topological structure of the real-world network graphs as well as to elucidate the impact of the degree and maximal clique size distribution of the vertices on the correlation between the centrality values and the maximal clique size observed for the vertices. The network graphs analyzed are briefly described as follows: (i) *Zachary's Karate Club* [11]: Social network of friendships (78 edges) between 34 members of a karate club at a US university in the 1970s;

(ii) *Dolphins' Social Network* [12]: An undirected social network of frequent associations (159 edges) between 62 dolphins in a community living off Doubtful Sound, New Zealand; (iii) *US Politics Books Network* [13]: Nodes represent a total of 105 books about US politics sold by the online bookseller Amazon.com. A total of 441 edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon; (iv) *Word Adjacencies Network* [14]: This is a word co-appearance network representing adjacencies of common adjective and noun in the novel "David Copperfield" by Charles Dickens. A total of 112 nodes represent the most commonly occurring adjectives and nouns in the book. A total of 425 edges connect any pair of words that occur in adjacent position in the text of the book; (v) *US College Football Network* [15]: Network represents the teams that played in the Fall 2000 season of the American Football games and their previous rivalry - nodes (115 nodes) are college teams and there is an edge (613 edges) between two nodes if and only if the corresponding teams have competed against each other earlier; (vi) *US Airports* 1997 *Network*: A network of 332 airports in the United States (as of year 1997) wherein the vertices are the airports and two airports are connected with an edge (a total of 2126 edges) if there is at least one direct flight between them in both the directions. Data for networks (i) through (v) can be obtained from http://www-personal.umich.edu/~mejn/netdata/. Data for network (vi) can be obtained from: http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm.



US College Football Network (115 nodes, 613 edges)

Dolphins' Social Network (62 nodes, 159 edges)

US Politics Books Network (105 nodes, 441 edges)

Karate Club Network (34 nodes, 78 edges)

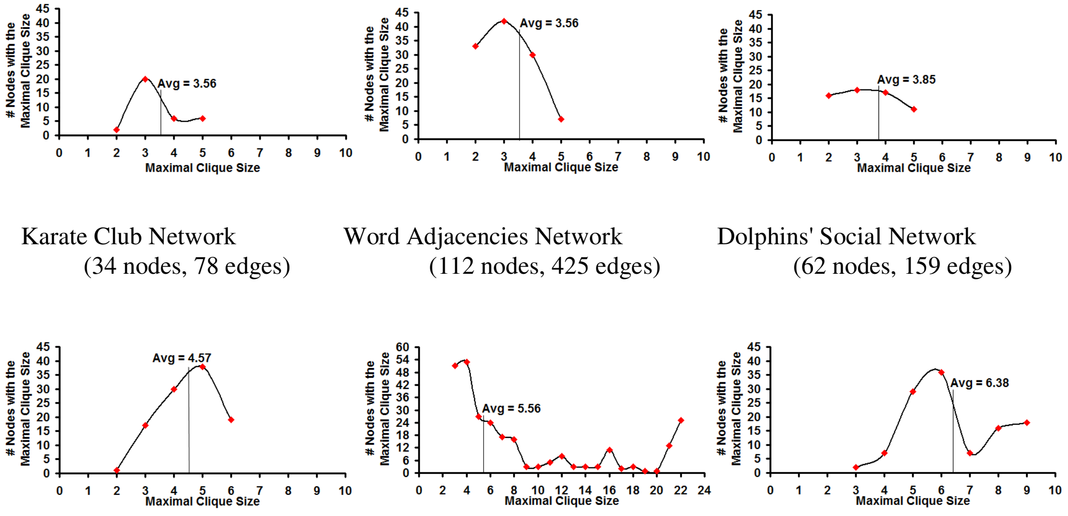
Word Adjacencies Network (112 nodes, 425 edges)

US Airports'97 Network (332 nodes, 2126 edges)

Figure 3. Distribution of Node Degrees (Probability Mass Function and Cumulative Distribution)

4.1. Degree Distribution of the Real-World Network Graphs

Figure 3 presents the degree distribution of the vertices in the six network graphs in the form of both the Probability Mass Function (the fraction of the vertices with a particular degree) and the Cumulative Distribution Function (the sum of the fractions of the vertices with degrees less than or equal to a certain value). We also compute the average node degree and the spectral radius degree ratio (ratio of the spectral radius and the average node degree); the spectral radius (bounded below by the average node degree and bounded above by the maximum node degree) is the largest eigen value of the adjacency matrix of the network graph, obtained as a result of computing the eigenvector centrality of the network graphs. The spectral radius degree ratio is a measure of the variation in the node degree with respect to the average node degree; the closer the ratio is to 1, the smaller the variations in the node degree and the degrees of the vertices are closer to the average node degree (characteristic of random graph networks). The farther is the ratio from 1, the larger the variations in the node degree (characteristic of scale-free networks). Figure 3 presents the degree distribution of the network graphs in the increasing order of their spectral radius ratio for node degree (1.01 to 3.23). The US College Football network exhibits minimal variations in the degree of its vertices (each team has more or less played against an equal number of other teams). The US Airports network exhibits maximum variation in the degree of its vertices (there are some hub airports from which there are flights to several other airports; whereas there are several airports with only fewer connections to other airports). In between these two extremes of networks, we have the other four network graphs, all of which have a spectral radius ratio for node degree around 1.4-1.7, indicating a moderate variation in the node degree (compared to the spectral radius ratios observed for the US College Football network and the US Airports network).



Karate Club Network (34 nodes, 78 edges)

Word Adjacencies Network (112 nodes, 425 edges)

Dolphins' Social Network (62 nodes, 159 edges)

US Politics Books Network (105 nodes, 441 edges)

US Airports 1997 Network (332 nodes, 2126 edges)

US College Football Network (115 nodes, 613 edges)

Figure 4. Distribution of Maximal Clique Size of the Vertices in Real-World Network Graphs

4.2. Maximal Clique Size Distribution of the Real-World Network Graphs

Figure 4 presents the distribution of the maximal clique size of the vertices for the six real-world network graphs, in the increasing order of the average value for the maximal clique size of the vertices. An interesting observation is that five of the six real-world network graphs exhibit a

Poisson-style distribution for the maximal clique size and the average value of the maximal clique size for the nodes is very close to the maximum value. The only real-world network that does not exhibit a Poisson-style distribution for the maximal clique size is the US Airports network whose distribution of the maximal clique size appears to be more of a scale-free (power-law) pattern with a long tail (wherein the average maximal clique size is 5.56, but there exists a significant number of nodes whose maximal clique size values are 21 and 22). We can also notice that the average value of the maximal clique size of the nodes is not proportional to the number of nodes in the network nor to the spectral radius ratio for node degree. This is evident from three of the six real-world networks with comparable number of nodes (Word Adjacency Network - 112 nodes, US Politics Books Network - 112 nodes and the US Football Network - 105 nodes) incurring significantly different average values for the maximal clique size (3.56, 4.57 and 6.38 respectively). Similarly, though the spectral radius ratio for node degree increases with increase in the scale-free nature of the networks, we do not observe any such pattern of increase or decrease for the maximal clique size; for example: the US Football Network, Word Adjacency Network and the US Airports Network have spectral radius ratio for node degree values of 1.01, 1.73 and 3.22 respectively; whereas, their average maximal clique size values are 6.38, 3.56 and 5.56 respectively (no pattern of increase or decrease with increase in the spectral radius ratio for node degree).

5. CORRELATION COEFFICIENT ANALYSIS: MAXIMAL CLIQUE SIZE VS. CLUSTERING COEFFICIENT

The clustering coefficient of a node is the ratio of the number of links between the neighbors of the node to that of the maximum possible number of links between the neighbors of the node [1]. If a node i has k_i neighbors and there are L_i links among these k_i neighbors, then the clustering coefficient for node i is: $C_i = \frac{L_i}{k_i(k_i - 1)/2}$. In this section, we examine whether the clustering coefficient of the nodes in the six real world network graphs is positively correlated to the maximal clique size of the nodes in these graphs. Our reasoning is that a clique comprises of links between any two of its constituent nodes; thus, the neighbors of a node in a clique are also connected with links among themselves. We wanted to examine whether or not this corresponds to links between any two neighbors of a node in the real world network graph itself.

If $\overline{X}$ and $\overline{Y}$ are the average values of the two metrics (say X and Y) observed for the vertices (IDs 1 to n, where n is the number of vertices) in the network, the formula used to compute the Correlation Coefficient between two metrics X and Y is given in equation (1), as follows:

$$CorrCoeff(X,Y) = \frac{\sum_{ID=1}^{n}(X[ID] - \overline{X}) * (Y[ID] - \overline{Y})}{\sqrt{\sum_{ID=1}^{N}(X[ID] - \overline{X})^2}\sqrt{\sum_{ID=1}^{N}(Y[ID] - \overline{Y})^2}} \quad (1)$$

Table 1. Correlation Coefficient between Maximal Clique Size and Clustering Coefficient

Network Index	*Network Name*	*Spectral Radius Ratio for Node Degree*	*Correlation Coefficient: Maximal Clique Size vs. Clustering Coefficient*
(vi)	US Airports 1997 Network	3.22	-0.47
(i)	Karate Club Network	1.46	-0.22
(ii)	Dolphins' Social Network	1.40	-0.17
(iv)	Word Adjacencies Network	1.73	-0.09
(iii)	US Politics Books Network	1.41	0.07
(v)	US College Football Network	1.01	0.69

Table 1 lists the correlation coefficient observed for the clustering coefficient and the maximal clique size of the nodes for the six real world network graphs (in the order of increasing values of the correlation coefficient), along with the spectral radius ratio for node degree observed for these networks. Contrary to our hypothesis, we observe the clustering coefficient of the nodes in four of the six real world network graphs to be poorly correlated to the maximal clique size of the nodes; the exceptions being the US College Football network (a random network graph) and the US Airports' 97 network (a scale-free network graph) exhibiting respectively moderate levels of positive and negative correlations between the clustering coefficient and the maximal clique size of the nodes. Hence, if at all a positive correlation is observed between the clustering coefficient and maximal clique size, it is most likely by chance. On the other hand, the correlation between the clustering coefficient and maximal clique size turns more negative with increase in the scale-free nature of the networks. For networks that have moderate values of the spectral radius ratio for node degree (that is the networks are neither scale-free nor random), there is pretty much no correlation between the clustering coefficient and maximal clique size of the nodes.

6. Correlation Coefficient Analysis: Maximal Clique Size vs. Node Degree

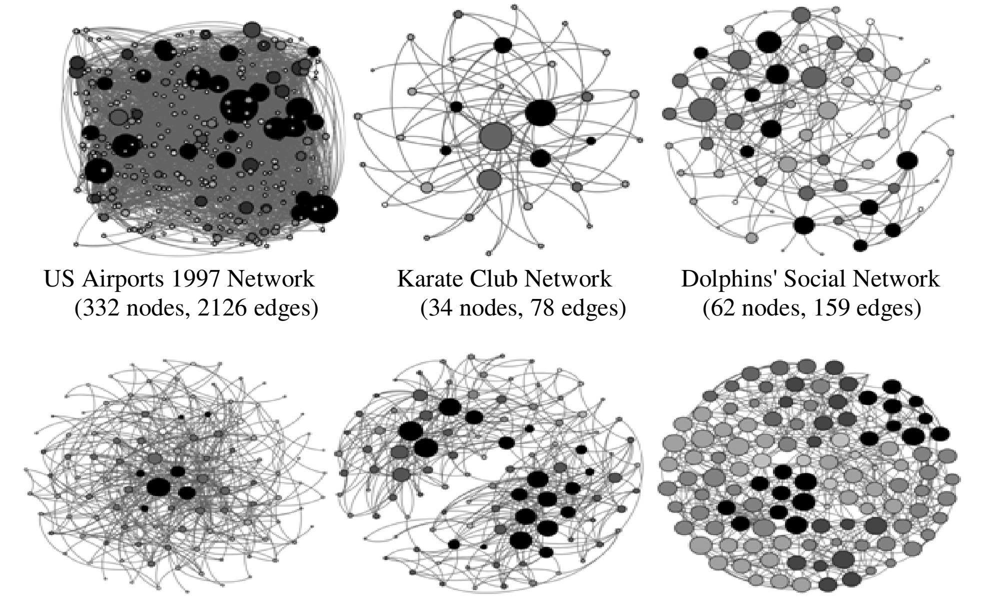
In this section, we present the results of correlation coefficient analysis conducted between node degree vis-a-vis the maximal size clique that each vertex is part of. The analysis has been conducted on the six real-world network graphs (mentioned in Section 4) with respect to the node degree and the maximal clique size measured for the vertices in these graphs. We implemented the exact algorithm to determine the maximal clique size for each of the vertices in a graph (see Figure 2). The visualization figures presented in the paper were obtained by porting the network data as well as the node degree/maximal clique size results to Gephi [16] and making appropriate changes to the settings in the latter. The layout algorithm chosen in Gephi for visualization is the Fruchterman Reingold algorithm [17] that presents the network in a circular format (like a globe).

Table 2. Correlation Coefficient between Maximal Clique Size and Node Degree

Network Index	*Network Name*	*Spectral Radius Ratio for Node Degree*	*Correlation Coefficient: Maximal Clique Size vs. Node Degree*
(vi)	US Airports 1997 Network	3.22	0.87
(i)	Karate Club Network	1.46	0.64
(ii)	Dolphins' Social Network	1.40	0.78
(iv)	Word Adjacencies Network	1.73	0.71
(iii)	US Politics Books Network	1.41	0.70
(v)	US College Football Network	1.01	0.32

Table 2 presents a correlation coefficient analysis of node degree and the maximal clique size observed for the vertices in each of the six real-world network graphs (in the decreasing order of the spectral radius ratio for node degree). As we can see in Table 2, in general, the correlation between the node degree and the maximal clique size increases as the spectral radius ratio for node degree increases. This implies, the more scale-free a real-world network is, the higher the correlation between the centrality value and the maximal clique size observed for a node. With several of the real-world networks being mostly scale-free, we expect these networks to exhibit a similar correlation to that observed in this paper. Also, since the correlation between the maximal clique size and node degree is the lowest (correlation coefficient of 0.31) for the US College Football Network (a random network), we conjecture that the stronger correlation (correlation coefficient of 0.7 or above) observed between these two metrics in the other five real-world network graphs is not merely by chance.

Figure 5 depicts the correlation observed for the node degree with that of the maximal clique size for the vertices in the real-world network graphs. In these figures, the node size is a measure of the node degree (the larger a node is, the larger is its degree); the node color is a measure of the maximal size clique the vertex is part of (the darker a node is, the larger is the size of the maximal clique for the node). We observe a high correlation between maximal clique size of nodes and nodes with a higher degree as well as located in a neighborhood of high degree nodes. That is, a node with high degree as well as located in a neighborhood of high degree vertices is more likely to be part of a maximal clique of larger size. In addition, as the networks get increasingly scale-free, nodes with high degree are more likely connected to other similar nodes with high degree (to facilitate an average path length that is almost independent of network size: characteristic of scale-free networks [1]) contributing to a positive correlation between degree-based centrality metrics and maximal clique size. We anticipate that as the networks become increasingly scale-free, the hubs (that facilitate shortest-path communication between any two nodes) are more likely to form the maximum clique for the entire network graph - contributing to higher levels of positive correlation between node degree and maximal clique size.



US Airports 1997 Network (332 nodes, 2126 edges)

Karate Club Network (34 nodes, 78 edges)

Dolphins' Social Network (62 nodes, 159 edges)

Word Adjacencies Network (112 nodes, 425 edges)

US Politics Books Network (105 nodes, 441 edges)

US College Football Network (115 nodes, 613 edges)

Figure 5. Correlation of Maximal Clique Size of the Vertices and Node Degree in the Real-World Network Graphs

7. Assortativity Index-Based Analysis: Maximal Clique Size And Node Degree

The assortativity index for a network graph with respect to a particular node-related metric is a measure of the association of nodes with similar values for the metric [1]. For example, the assortativity index of a graph with respect to node degree is a measure of the association of higher degree nodes with other high degree nodes as well as the association of nodes of lower degree nodes with other lower degree nodes. In this section, we analyze the assortativity index of the six real-world network graphs with respect to the maximal clique size and node degree, and examine the nature of association between nodes having higher values for each of these two metrics. If *m* is the node-related metric of interest, then the assortativity index of the network graph with respect to *m* is evaluated as the correlation coefficient of the values (with respect to metric *m*) for the end nodes of the edges in the graph. Consider a network graph of *n* nodes and set of undirected (bi-directional) edges *E*; let *m*[1], *m*[2], ..., *m*[*n*] be the values for nodes 1, 2, ...,*n* with respect to metric *m* and $\overline{m}$ be the average value of the metric, the assortativity index with respect to metric *m* is given by equation (2).

$$AssortativityIndex(m) = \frac{\sum_{(i,j)\in E} (m[i]-\overline{m})*(m[j]-\overline{m})}{\sqrt{\sum_{(i,j)\in E} (m[i]-\overline{m})^2}\sqrt{\sum_{(i,j)\in E} (m[j]-\overline{m})^2}} \qquad (2)$$

Positive values for the assortativity index with respect to a metric indicates that the network exhibits assortativity with respect to the metric (nodes with higher values for the metric are more likely to be connected to nodes with higher values for the metric and vice-versa); negative values for the assortativity index indicates the network exhibits disassortativity (nodes with lower values for the metric are more likely to be connected to nodes with higher values for the metric and vice-versa); assortativity index values close to 0 indicates the network is more neutral with respect to the metric (i.e., the values for the end nodes of the edges with respect to the metric do not exhibit any correlation).

Table 3 lists the assortativity index values for the maximal clique size and degree of the vertices for the six real-world network graphs, along with their spectral radius ratio for node degree. We observe the assortativity index (with respect to the maximal clique size) for all the six network graphs to be positive and the assortativity index value for the maximal clique size increases with increase in the level of randomness in the network, indicating that the association of nodes of a particular maximal clique size with other nodes that are also of the same maximal clique size is more by chance. On the other hand, we observe the assortativity index (with respect to the node degree) for five of the six network graphs (i.e., all network graphs, except the US Football Network that exhibits the characteristic of random graphs) to be negative and the assortativity index values for the node degree gets more negative with increase in the scale-free nature of the network, indicating high degree nodes are more likely to be associated with low degree nodes (especially with increase in the spectral radius ratio for node degree). Finally, to confirm our earlier observation of a positive correlation between maximal clique size of the vertices and node degree, we observe in Table 3 that the six-real world networks could be listed in an identical order, in the increasing order of the Assortativity Index of the network graphs with respect to both maximal clique size of the vertices and node degree.

Table 3. Assortativity Index of the Real-World Network Graphs based on Maximal Clique Size of the Vertices and Node Degree

Network Index	*Network Name*	*Spectral Radius Ratio for Node Degree*	*Assortativity Index for Maximal Clique Size*	*Assortativity Index for Node Degree*
(i)	Zachary's Karate Network	1.46	0.13	-0.48
(vi)	US Airports 1997 Network	3.22	0.17	-0.21
(iv)	Word Adjacencies Network	1.73	0.20	-0.09
(iii)	US Politics Books Network	1.41	0.20	-0.04
(ii)	Dolphins' Social Network	1.40	0.23	-0.02
(v)	US College Football Network	1.01	0.59	0.19

8. CONCLUSIONS

The observation of Poisson-style distribution for maximal clique size of the vertices in real-world network graphs irrespective of the number of nodes and the degree distribution of the vertices is an interesting observation that has not been hitherto reported in the literature. We conjecture the distribution for the maximal clique size of the vertices to transform from Poisson to Power-law distribution in networks that are highly scale-free (as observed in the case of the US Airports'97 Network). With the problem of determining maximal clique sizes for individual vertices being computationally time consuming, our approach taken in this paper to study the correlation between maximal clique sizes vis-a-vis node degree and clustering coefficient could be the first step in identifying correlation between cliques/clique size in real-world network graphs and one or more computationally-light node-based network metrics that can be quickly determined and henceforth appropriate inferences can be made about a ranking of the vertices with respect to maximal clique size. The approach taken to first to find the correlation coefficient between the two metrics of interest (like node degree and maximal clique size of the vertices) in the individual network graphs and then ranking the network graphs in the increasing order of the Assortativity Index of the graphs with respect to each of the two metrics (an identical or close to identical listing of the network graphs with respect to the each of the two metrics vindicates the positive correlation observed between the two metrics based on correlation coefficient analysis). Such an approach for correlation study between two node-based metrics is a unique approach that has been so far not presented in the literature. We observe node degree to show promising positive correlations to that of maximal clique sizes of the individual vertices, especially as the networks get increasingly scale-free; this observation could form the basis of future research for analysis of maximal clique size of the vertices in complex real-world network graphs and the correlations of the maximal clique size of the vertices with other computationally-light metrics for complex network analysis.

REFERENCES

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

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to Algorithms, 3rd ed., MIT Press, July 2009.

[3] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society," Nature, vol. 435, pp. 814-818, 2005.

[4] S. Fortunato, "Community Detection in Graphs," Physics Reports, vol. 486, pp. 75-174, 2010.

[5] S. Sadi, S. Oguducu and A. S. Uyar, "An Efficient Community Detection Method using Parallel Clique-Finding Ants," Proceedings of IEEE Congress on Evolutionary Computation, pp. 1-7, July 2010.

[6] B. Pattabiraman, M. A. Patwary, A. H. Gebremedhin, W-K. Liao and A. Choudhary, "Fast Problems for the Maximum Clique Problem on Massive sparse Graphs," Proceedings of the 10th International Workshop on Algorithms and Models for the Web Graph: Lecture Notes in Computer Science, vol. 8305, pp. 156-169, 2013.

[7] R. Carraghan and P. Pardalos, "An Exact Algorithm for the Maximum Clique Problem," Operations Research Letters, vol. 9, no. 6, pp. 375-382, 1990.

[8] P. R. J. Ostergard, "A Fast Algorithm for the Maximum Clique Problem," Discrete Applied Mathematics, vol. 120, no. 1-3, pp. 197-207, 2002.

[9] R. A. Rossi, D. F. Gleich and M. A., Patwary, "Fast Maximum Clique Algorithms for Large Graphs," Proceedings of the 23rd International Conference on World Wide Web Companion, pp. 365-366, April 2014.

[10] E. Tomita and T. Seki, "An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique," Proceedings of the 4th International Conference on Discrete Mathematics and Theoretical Computer Science, pp. 278-289, 2003.

[11] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," Journal of Anthropological Research, vol. 33, no. 4, pp. 452-473, 1977.

[12] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-lasting Associations," Behavioral Ecology and Sociobiology, vol. 54, pp. 396-405, 2003.

[13] V. Krebs, http://www.orgnet.com/divided.html, last accessed: March 22, 2015.

[14] M. Newman, "Finding Community Structure in Networks using the Eigenvectors of Matrices," Physics Review, E 74, 036104, September 2006.

[15] M. Girvan and M. Newman, "Community Structure in Social and Biological Networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 19, no. 12, pp. 7821-7826, April 2002.

[16] K. Cherven, Network Graph Analysis and Visualization with Gephi, 1st ed., Packt Publishing, Birmingham, UK, September 2013.

[17] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-Directed Placement," Software: Practice and Experience, vol. 21, no. 11, pp. 1129-1164, November 1991.

INTENTIONAL BLANK

LIVE TILES IN MAINFRAME

Robin Tommy, Ullas Ravi, Jobin Luke, Aswathy S Krishna,
Fathima Thasneem, Girija Subramanian

Tata Consultancy Services, Trivandrum

ABSTRACT

This paper proposes the usage of Live Tiles Usability Experience in Mainframes. The UI experience was brought into the mainframe CICS maps. The map related parameters and attributes are used for achieving the live tiles display and enhancing the user experience. The dynamic tiles UI is provided with location based interfacing on the CICS map. The concept of Temporary storage Queue (TSQ) is implemented in order to display the information and the concept of cursor positioning is also used to make the application user friendly. This paper also addresses how Live Tiles will improve the user interface in Mainframes. Live tile is implemented using COBOL, DB2, CICS in mainframe.

KEYWORDS

Live Tiles, Mainframes, Cursor positioning, CICS, TSQ.

1. INTRODUCTION

Mainframes is the oldest technology, which has been used mainly for high security based applications [1]. When we think of mainframes, the first thing that comes to our mind is the monotonous black screen and the not so user-friendly interface [2] so, over the years mainframes was seen as a back end application.

Since there isn't any attractive interface like Windows in Mainframes, implementing a concept like Live Tiles [3] could be alien to many of us. Live Tiles, which became famous through Windows, are the colourful blocks which help us in launching applications. So they are called as the shortcuts to the applications and we can click on them to open any particular application. The tiles are so much more than simple icons. They give us an at-a-glance view of a range of useful information, weather and news headlines are the few which come to our mind.

Live Tiles provides a dynamic compressed view of the information. The colour and presentation of tiles provides a user view of data in a different level. Different types of information are presented in a more user experience methodology providing a faster access to the important data for the user. The live tiles in mainframes define a new UX design in CICS maps which takes the mainframe maps to a different usability model. The important content for the user is personalized and brought on to the CICS mainframe screen with a dynamic information capability [4]. Live tiles uses CICS concepts such as cursor positioning, and pseudo conversational programming [6] [7].Detailed information of the data shown on the tile will be displayed with the help of Temporary Storage Queue (TSQ) [5].

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 15–19, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50902

2. IMPLEMENTATION

Customer Information Control System (CICS) is the user interface used and in order to display data TSQ [4] is used. Temporary storage is the primary CICS facility for storing data that must be available to multiple transactions. Data items in temporary storage are kept in temporary storage queues. Temporary Storage Queue is a feature that is provided by the Temporary Storage Control Program (TSP). A TSQ [5] is a queue of records that can be created, read and deleted by different tasks or programs in the same CICS region. The items can be retrieved by the originating task, or by any other task, by using the symbolic name assigned to the temporary storage queue.

A temporary storage queue containing multiple items can be thought of as a small data set. Specific items (logical records) in a queue are referred to by relative position numbers.

The records in TSQ remains accessible until the entire TSQ is explicitly deleted. The records in TSQ can be read sequentially or directly.

Next comes the cursor positioning [6], by clicking on the tile, the corresponding application should be opened. EIBCPOSN is used to achieve the same.

Range = (No of rows - 1) * 80 + (no of columns -1)

The above equation [7] is used to calculate the range for which the cursor positioning should work, thereby connecting the tile to its respective application. Combining the entire mainframe CICS, map and cursor positioning, live tiles is implemented with a drill down approach. Each tile has different data with dynamic content management.

3. RESULTS

On a daily basis, a TO-DO list along with a CALENDER and the LATEST NEWS update are the things we want to have a track of. We have implemented these features on Mainframes as LIVE TILES. (Figure 1)

The back ground process is listed as follows

- Check the date and time condition.
- Use a Cursor to fetch data from DB2.
- Use TSQ to store the data.
- Display the required information on the tiles.
- Cursor positioning

Highlighted dates in the Calendar (figure 1) tells that certain events are scheduled on those dates. By clicking on that date, the entire day's schedule can be viewed. The current date in “RED” has its latest three scheduled events displayed on the TO-DO(What’s Next) tile (figure 2).

In the NEWS update tile, the latest three news based on the login time can be viewed. By clicking anywhere on the tile the entire news headlines of that day are displayed. (Figure 3)

Following is the statistical results of using the live tiles.

1. Time taken to access the relevant data on the CICS map before and after live tiles. (Chart 1)

2. Average amount of information (maps) processed for getting the same information before and after using live tiles. (Chart 2)

Here we tried accessing the TO-DO list map before implementing the live tiles which needed a transition through 4 maps to reach the relevant data but after the live tiles implementation the information was made available on the first map.

Chart 1

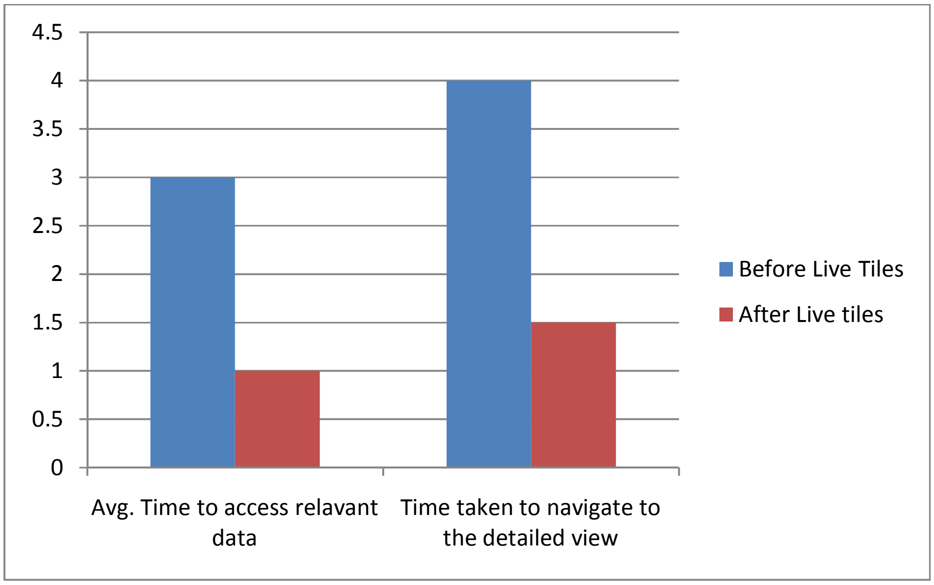
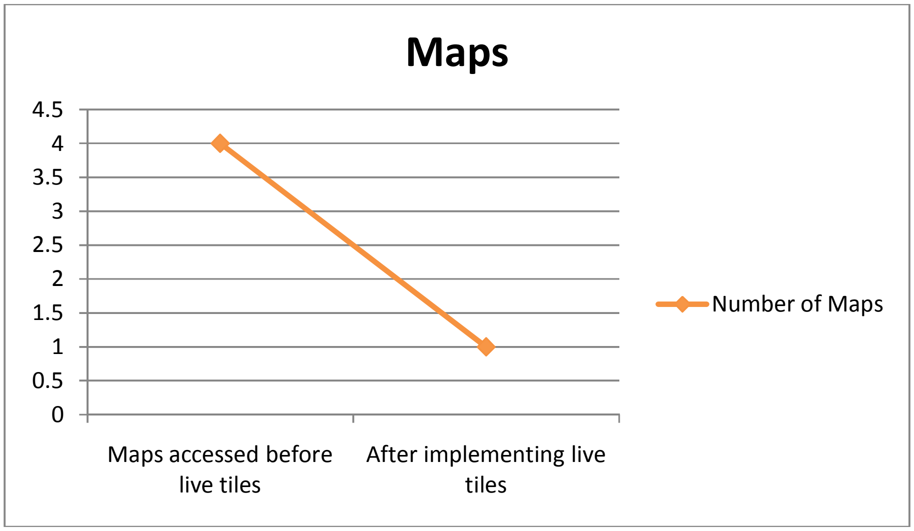


Chart 2

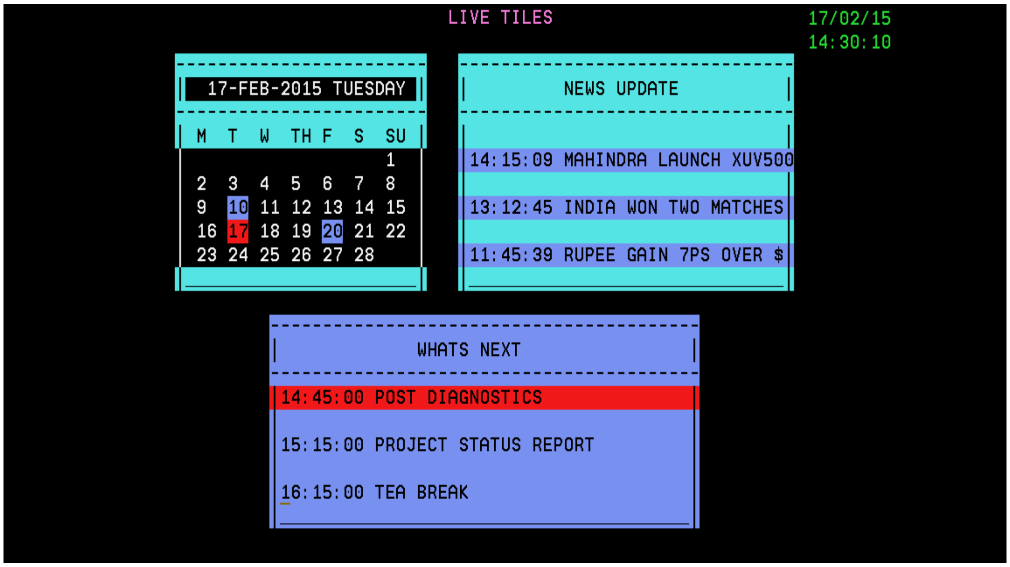


Thus the results show that the implementation of live tiles improved the efficiency of the UI in mainframes and provides the relevant information much earlier to the user.

3. Conclusion

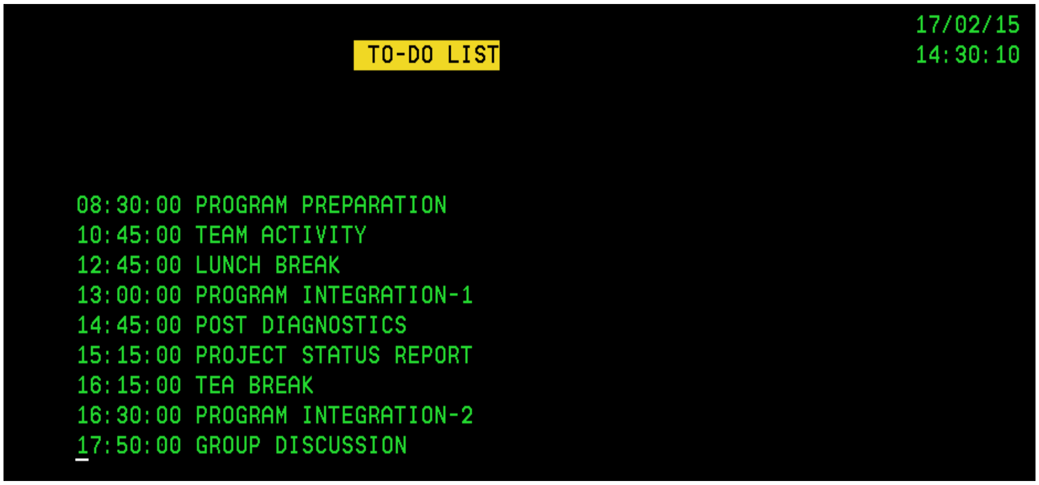
Tiles are nothing but a small piece of colourful rectangular blocks, which can be incorporated in mainframes as seen above, with live updates on it. Live Tiles in mainframes will help in improving the user interface as updated information can be directly viewed from the start screen/Home screen in a colourful way. This provides mainframe a different user experience with faster access to the relevant information.

Figures



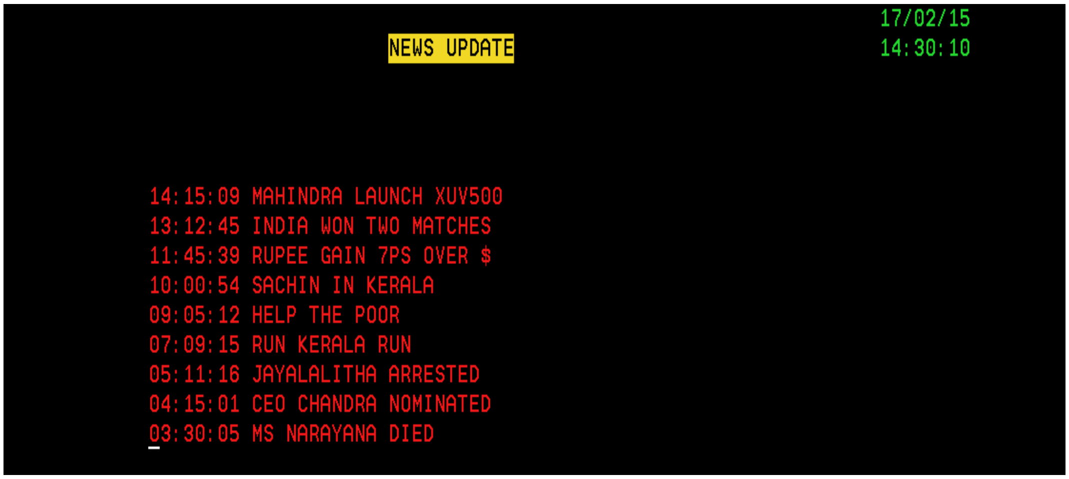
```
                         LIVE TILES                         17/02/15
                                                            14:30:10
 17-FEB-2015 TUESDAY           NEWS UPDATE

 M  T  W  TH F  S  SU
                   1           14:15:09 MAHINDRA LAUNCH XUV500
 2  3  4  5  6  7  8
 9  10 11 12 13 14 15          13:12:45 INDIA WON TWO MATCHES
 16 17 18 19 20 21 22
 23 24 25 26 27 28             11:45:39 RUPEE GAIN 7PS OVER $

                    WHATS NEXT
        14:45:00 POST DIAGNOSTICS

        15:15:00 PROJECT STATUS REPORT

        16:15:00 TEA BREAK
```

Figure 1: Live tiles home page



```
                                                   17/02/15
                     TO-DO LIST                    14:30:10



    08:30:00 PROGRAM PREPARATION
    10:45:00 TEAM ACTIVITY
    12:45:00 LUNCH BREAK
    13:00:00 PROGRAM INTEGRATION-1
    14:45:00 POST DIAGNOSTICS
    15:15:00 PROJECT STATUS REPORT
    16:15:00 TEA BREAK
    16:30:00 PROGRAM INTEGRATION-2
    17:50:00 GROUP DISCUSSION
```

Figure 2: TO-DO List Screen



```
                                                              17/02/15
                                NEWS UPDATE                   14:30:10



            14:15:09 MAHINDRA LAUNCH XUV500
            13:12:45 INDIA WON TWO MATCHES
            11:45:39 RUPEE GAIN 7PS OVER $
            10:00:54 SACHIN IN KERALA
            09:05:12 HELP THE POOR
            07:09:15 RUN KERALA RUN
            05:11:16 JAYALALITHA ARRESTED
            04:15:01 CEO CHANDRA NOMINATED
            03:30:05 MS NARAYANA DIED
```

Figure 3: News update screen

REFERENCES

[1] Paulson, Linda Dailey, "Mainframes, Cobol still popular" IT Professional, IEEE, vol.3, no.5, pp.12,14, Sep/Oct 2001 doi: 10.1109/6294.952975

[2] Effective GUI Implementation in Mainframes – Windows based, IJRDE, Vol.1:Issue.2,October November 2012 pp-11-17

[3] https://msdn.microsoft.com/en-us/library/windows/apps/dn468032.aspx

[4] Implementation of DashBoard in Mainframes For Business Analysis, IJRDE, Vol:Issue.2,October-November 2012pp-18-27.

[5] Implementation of Word prediction in Mainframes, IJRDE, Vol.1: Issue.3, February-March 2013 pp 55-62.

[6] Data Modelling Based on Usage Analytics in Mainframe CICS Applications, IJRDE, Vol.1: Issue.2, October-November 2012pp-37-45.

[7] IBM Mainframe Handbook by Alexis Leon.

INTENTIONAL BLANK

USING ADAPTIVE SERVER ACTIVATION/DEACTIVATION FOR LOAD BALANCING IN CLOUD-BASED CONTENT DELIVERY NETWORKS

Darshna Dalvadi[1] and Dr. Keyur Shah[2]

[1]School of Computer Studies, Ahmedabad University, Ahmedabad , India.
darshnadalvadi@gmail.com
[2] Head of the Department, L.D.R.P Institute of Technology and Research, Kadi Sarv Vidhyalaya, Gandhinagar, India.
profkeyur@gmail.com

ABSTRACT

Content Delivery Networks have been widely used for many years to serve million of users. Lately, many of these networks are migrating to the cloud for its numerous advantages such as lower costs, availability and dynamism in resource provisioning for obtaining overall increased performance. This paper introduces a new approach towards load balancing as well as lower response time(limited latency) in cloud-based content delivery networks allowing for providing overall improved performance within the tradeoff between power consumption and quality of experience (QoE). New proposed adaptive server activation/ deactivation model aims towards switching off unutilized servers at the data center to reduce the power consumption and also to use available resources at maximum efficiency while maintaining its performance. Latency can be lowered by only shifting the load off a datacenter when it is almost fully loaded.

KEYWORDS

Cloud Computing, Content Delivery, Load Balancing, Power reduction, Resource Management, Adaptive Self Control, Performance Modeling.

1. INTRODUCTION

This CDN is aimed towards serving end-users a large fraction of the Internet content like text, graphics and scripts, media files, software, documents, e-commerce, portals, live streaming media, on-demand streaming media, and social networks with good QoE. CDNs allow for limited latency by distributing content at servers across different region or worldwide and serving users by considering their geographical location, Origin of web page and type of requested data. Many distributed servers are doing jobs on behalf of original content server. Content is served by nearest sever with maintaining load balancing. Recently, many CDN providers started integrating their networks with the cloud as the cloud provides numerous advantages for both CDN users and providers. Using cloud CDN can be easily extended as resources could be rent from the cloud provider on demand. Also users will be benefited as they will no longer need to install physical storage devices to be part of the CDN, and will only pay for the content usage and content transfer which reduces the operating costs rapidly. However , there are two issues needed to be considered in cloud based CDN: 1) load balancing 2) network latency. Load balancing among

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 21–25, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50903

different data centers in the cloud need to be done in parallel with locality awareness. This provides least delay in transferring content. A good performance metric should be considered while communication is latency which can be minimized by caching contents at different servers instead of serving from origin server. This study introduces an approach to load balancing in cloud based CDNs by offloading data centers when it is loaded beyond certain threshold and also maintaining delay bounds instead of equally distributing workload among servers so that data can be stored on data centers nearest to users who usually request it. In previous work a new algorithm was introduced to adapt the number of active servers in any datacenter accordingly the amount of offered network load at any time by using multi-level, parallel hysteresis threshold algorithm[1][2]. Results have shown that by applying this algorithm to any multiple server data center, the delay experienced by users observe an almost constant behavior over a vast range of offered load. This implies that if the Load increases up to a certain limit at a data center, the latency of content delivery can be limited and adapted to the users' SLA accordingly. By using this property provided by the algorithm, user requests will always be routed to the nearest data center storing the requested data until the load on the data center reaches a certain threshold, afterward requests are routed to the nearest under-loaded data center.

2. ADAPTIVE SERVER ACTIVATION/DEACTIVATION MODEL

A. Model without considering Server Activation Overhead

A load balancing method used in this paper is called "Multiple Parallel Hysteresis Model". This method adjusts the number of active servers according to offered load by switching off and on only at a certain threshold. There are two different thresholds for activation and deactivation to take place. Consider M/M/n/s queue model. In this extended notation C/B/n/s; C=M indicates the arrival rate of requests as Markovian (i.e., negative-exponentially distributed inter arrivals times), B=M indicates negative-exponentially distributed service time*s, n denotes the total number of* servers, and denotes the finite number of buffers for requests (frames). The parameters λ and μ indicate the arrival and service rates, respectively. The load factor

$$A = {}^{\lambda}/_{\mu} \tag{1}$$

indicates the average number of occupied servers as A<n. In a finite buffer system further request arriving is to be rejected when buffers are full. Each system state is represented by a pair(x,z)where x is the number of active servers and z is the number of buffered frames. Now, instead of activating server at each arrival of request, requests are buffered until certain threshold has reached, then only activation is done. Activation thresholds are determined by the number of buffered frames, namely $w^{(1)}$, $w^{(2)}$, $w^{(3)}$, …$w^{(n-1)}$, where $w^{(i)}$ is the width of the i^{th} hysteresis and $w^{i} = w^{(i)} - w^{(i-1)} \geq 0$ indicates the increase in buffered units for x=i until the next server is activated, i=1,2,…,n-1 where $w^{(i)}=w^{1}+\ldots+w^{i}$ and $w^{(0)}=0$.These hysteresis widths can be adjusted in a such way to meet the SLA's within the power reduction/performance tradeoff scenario. Deactivation policy works similar to any M/M/n queue where deactivations take place only when a server becomes idle and no buffered requests remain. The stationary state probabilities $P(x,z)$ which can be determined by selecting activation and deactivation threshold is discussed in [3]. Applying this model increases probability of selecting hysteresis state where amount of load is like X or X+1 servers are active, and reduces probability of all other states. Average delay experienced by delayed customers in this model is measured using Little's Theorem according to the following relations:

- Average number of delayed arrivals

$$L = \sum_{x=1}^{n} \sum_{z=1}^{s} P(x,z) \tag{2}$$

- Probability of buffering an arrival

$$W = 1 - P(0{,}0) - P(n,s) \qquad (3)$$

where arrivals to the system at state (*0*,*0*) are served immediately and at state (*n*,*s*) are dropped.

- Mean waiting time of buffered arrivals

$$E[Tw|Tw > 0] = \frac{L}{\lambda W} \qquad (4)$$

- Probability of lost arrivals

$$B = P(n,s) \qquad (5)$$

B. Model with Server Activation Overhead

Practically, deactivated servers take some time to become activate depending on their current deactivation mode. If node is switched off (disabled power supply) so it has to be booted from starting, which takes more time. If server is set in a sleep mode then it takes lower time to be activated. If we consider this time in estimating average delay in above discussed scenario then calculation is as follows: when the sum of activate servers and number of buffered frames in all the states with lower number of busy servers, a server is triggered to become activate. The rate at which a server is triggered to be activated is α, then average overhead time for an idle server to become active is $^{1}/_{a}$ add this delay factor in mean waiting time.

3. LOAD BALANCING STRATEGY

Implementation of the load balancing strategy explained in this paper shows that, the load adaptive model applied on each datacenter of CDN allows the data center to have almost fixed delays over a vast range of loads. This allows for increasing the load on the data center up to utilization factor of 95% and more, so requests are always routed to the nearest data center without lowering the quality of offered service even if the data center is highly loaded. Certain assumptions are as follows:

- Total N number of data centres are there
- Each data center has n_i servers
- Each data centre has offered load

$$A_i = \frac{\lambda_i}{\mu_i} \qquad (6)$$

The algorithm steps are as follows:

1. Determine the maximum load that could be handled by each data center

$$A_{(max,i)} = [function(n_i)|t_w < t_{SLA}] \qquad (7)$$

where $A_{(max,i)}$ is a function of the number of servers n_i in each data center *i* and the maximum tolerable delay according to the users' SLA t_{SLA}.

2. Determine the load margin

$$\Delta A(i) = A_i - A_{(max,i)} \qquad (8)$$

If $\Delta A(i) > 0$: Data center i is overloaded and the extra load $\Delta A(i)$ needs to be shifted to another data center. If $\Delta A(i) \leq 0$: Data center i can still handle extra load equal to $\Delta A(i)$ without affecting its performance.

3. For DCs whose $\Delta A(i) > 0$, shift this amount of load to the nearest DC who can accommodate this load shift, fully or partially.

4. Repeat the above three steps until no more load shifting is necessary.

The question is how to select the nearest datacenter which can handle the load shift fully or partially two approaches are there:

1. Centralized: Each datacenter sends calculated ΔA to the center entity(cloud manager) and this entity decides where to shift the load.

2. Decenterlized: Each datacenter broadcast their ΔA to other datacenters and overloaded datacenter decides where to shift the load. Here, as long as the load is within the stable delay region, all the datacenters works in a self-controlled manner and needs no longer external control mechanism. The activation-deactivation model is sufficient to handle stable delay within the performance and power reduction tread off.

4. Conclusion

Paper introduced a new model for load balancing in cloud based CDN. Model employs following approaches:

1. Apply multiple parallel hysteresis model on each datacenter for providing limited delay.

2. Model attempts to adapt number of active servers accordingly the offered load which causes delay to be constant over a vast range of load. This allows routing users' requests to the nearest data center while maintaining the maximum load it can handle.

3. It also does load balancing by shifting any extra load from overloaded datacenter to the nearest neighbors which can handle this load coming from other data center without affecting the SLA's of users.

The model forces the data center to adapt the number of active servers to the offered load which causes the delay value to be constant over a vast range of values. This allows routing users' requests to the nearest data center to guarantee limited delays even if the data center was highly loaded. Also shifting of any extra load from overloaded servers to other under-loaded ones could be done without affecting their performance or affecting the users' service level agreements. The proposed approach balances the load between different cloud-based data centers while maintaining low delays. The model parameters can be adapted such that the additional delay resulting from buffering still meets SLA requirements.

REFERENCES

[1] Kuehn, P.J., Mashaly, M.: “Modeling and Performance Evaluation of Self-Adapting Algorithms for the Optimization of Power-Saving Operation Modes”, Proc. 1st European Teletraffic Seminar (ETS), Poznan, Poland, February 14-16,2011

[2] Kuehn, P.J.: “Systematic Classification of Self-Adapting Algorithms for Power-Saving Operation Modes of ICT Systems”, submitted in contribution to the 2nd Int. Conf. on Energy-Efficient Computing and Networking (e-Energy 2011), New York, USA, May 30 - June 1, 2011

[3] Kuehn, P.J., Mashaly, M.: “Performance of Self-Adapting Power-saving Algorithms for ICT Systems”, Forthcoming paper (submitted)

AUTHOR

Darshna Dalvadi received Master of Computer Applications from H.N.G University, in 2007. Working as a lecturer at Ahmedabad University and pursuing Ph.D. from CUSHAH University.

Dr. Keyur shah is working has a Head of the Department at L.D.R.P, Institute of Technology and Research, Gandhinagar. He is having 12 years of experience in academic field also guiding no. of Ph.D research scholars.

INTENTIONAL BLANK

AUTOMATIC COMPUTATION OF CDR USING FUZZY CLUSTERING TECHNIQUES

Thresiamma Devasia[1], Poulose Jacob[2] and Tessamma Thomas[3]

[1]Department of Computer Science, Assumption College Changanacherry,
Kerala, India
cherukusumam@gmail.com
[2]Department of Computer Science,
Cochin University of Science and Technology, Kerala, India
kpj0101@gmail.com
[3]Department of Electronics,
Cochin University of Science and Technology, Kerala, India
tess@cusat.ac.in

ABSTRACT

Eye disease identification techniques are highly important in the field of ophthalmology. A vertical Cup-to-Disc Ratio which is the ratio of the vertical diameter of the optic cup to that of the optic disc, of the fundus eye image is one of the important signs of glaucoma. This paper presents an automated method for the extraction of optic disc and optic cup using Fuzzy C Means clustering technique. The validity of this new method has been tested on 454 colour fundus images from three different publicly available databases DRION, DIARATDB0 and DIARETDB1 and, images from an ophthalmologist. The average success rate of optic disc and optic cup segmentation is 94.26percentage. The scatter plot depicts high positive correlation between clinical CDR and the CDR obtained using the new method. The result of the system seems to be promising and useful for clinical work.

KEYWORDS

Fundus image, optic disc, optic cup, Cup-to-Disc Ratio

1. INTRODUCTION

The fundus images are used for diagnosis by trained clinicians to check for any abnormalities or any change in the retina. Important anatomical structures captured in a fundus image are blood vessels, optic cup (OC), optic disc (OD) and macula for a normal retina. An image of a diseased retina may also contain many visible symptoms of the eye-disease. In a healthy retinal image the OD usually appears as a circular shaped bright yellowish object which is partly covered with vessels. The optic cup is the cupping of the optic nerve and that means the size of the depression in the middle of the nerve when viewed from the front of the eye. When there is damage to the optic nerve, the cupping increases. Changes in the OD and OC can indicate the presence, current state and progression of glaucoma [1][2]. An efficient segmentation of OD and OC is essential to diagnose various stages of certain diseases like glaucoma. The automatic computation of Cup-to-Disc Ratio (CDR) helps the ophthalmologist to do the screening and detection of glaucoma easily. In this paper, the vertical CDR is calculated by using fundus photograph where vertical CDR, is an important indicator of glaucoma [3]. Since the colour fundus images provide early signs of certain diseases such as diabetes, glaucoma etc., colour fundus images are used to track the eye

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 27–40, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50904

diseases by the ophthalmologists. Figure1 shows the important features of a retinal colour fundus image.

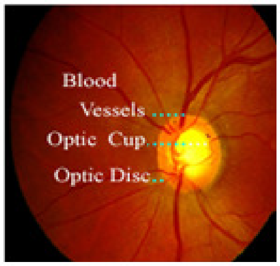


Figure 1. Colour Fundus Image

This paper is organized as follows:

Section II presents a brief survey of existing literature. Section III describes the materials used for the present work. A new algorithm to efficiently extract optic disc and optic cup in ocular fundus images is given in Section IV. The results are presented in Section V, and Conclusions are given in the final Section VI.

2. LITERATURE SURVEY

The Active Shape Model (ASM) based optical disk detection is implemented by Huiqi et al.[4]. The initialization of the parameters for this model is based on Principal Component Analysis technique. The faster convergence rate and the robustness of the technique are proved by experimental results. Huajun Ying et al. [5] designed a fractal-based automatic localization and segmentation of optic disc in retinal images. K. Shekar [6] developed a method for OD segmentation using Hussain, A.R. et al. [7] proposed a method for optic nerve head segmentation using genetic active contours. Zhuo Zhang et al. [8] designed a convex hull based neuro-retinal optic cup ellipse optimization technique. Wong, D.W.K. et al. [9] developed SVM-based model optic cup detection for glaucoma detection using the cup to disc ratio in retinal fundus images. Joshi G.D. et al. [10] developed vessel bend-based cup segmentation in retinal images. Shijian Lu et al. [11] proposed a background elimination method for the automatic detection of OD. Morphological operations were used for locating the optic disc in retinal images by Angel Suero et al. [12].

In this paper, a new algorithm based on Fuzzy C-Means Clustering (FCM) technique combined with thresholding, is used for OD and OC extraction. This new method, firstly, extracts the OD and OC of the colour fundus image and computes the vertical CDR automatically. This is an efficient method for the automatic screening of colour fundus image for CDR computation.

3. MATERIALS AND METHODS

The fundus images used in these experiments are taken from publicly available databases DRION, DIARATDB0 and DIARETDB1 and, images from Giridhar Eye Institute, Kochi, Kerala. The CDRs obtained from an ophthalmologist is used as ground truth for the evaluation.

4. DEVELOPED ALGORITHM

The new approach is composed of four steps. The channels of the colour retinal are separated. The blood vessels are removed, applying the contrast adjustment to enhance the low contrast image image. The Fuzzy C Means combined with thresholding is applied on the red channel of the input image for the extraction of the OD and the same technique is applied on the green channel of the input image for the extraction of OC. The CDR is computed using the ratio of vertical diameter of OC and OD.

4.1 Preprocessing

The preprocessing step excludes variations due to image acquisition, such as inhomogeneous illumination. In preprocessing, techniques such as morphological operations and contrast enhancement are applied on the input image [13]. The following sections include different preprocessing operations used in this paper.

4.1.1 Preprocessing steps for Optic Disc Extraction

4.1.1.1 Selection of Red Channel

From the previous studies it is shown that even though the green component of an RGB retinography is the one with highest contrast, the OD is often present in the red field as a well-defined white shape, brighter than the surrounding area [14]. Therefore the red channel of the RGB colour images is used for the extraction of OD regions in the retinal fundus images.

4.1.1.2 Removal of Blood Vessels

Since blood vessels within the OD are strong distracters, they should be erased from the image beforehand. In this method a morphological closing operation is performed on the red channel. The dilation operation first removes the blood vessels and then the erosion operation approximately restores the boundaries to their former position.

$$\text{Closing} \quad : C(A,B) = A \bullet B = E(D(A,-B),-B) \qquad (1)$$

where A is the red channel of the input image and B is a 10x10 symmetrical disc structuring element, to remove the blood vessels[15]. C is the resultant vessel free, smoothed output image.

4.1.2 Preprocessing steps for Optic Cup Extraction

4.1.2.1 Selection of Green Channel

The green channel has low contrast variation which gives more differentiation between the blood vessel and OC. The green channel, therefore, is selected for the extraction of the OC of the retinal image.

4.1.2.2 Removal of Blood Vessels

Blood vessels in the green channel were removed using a morphological closing procedure,

$$I2(I,B) = A \bullet B = E(D(I,-B),-B) \qquad (2)$$

where I is the green channel of the input image and B is an 8x8 symmetrical disc structuring element, to remove the blood vessels[13]. I2 is the smoothed, vessel free output image. Figure 2 shows the preprocessing operations on the input image.

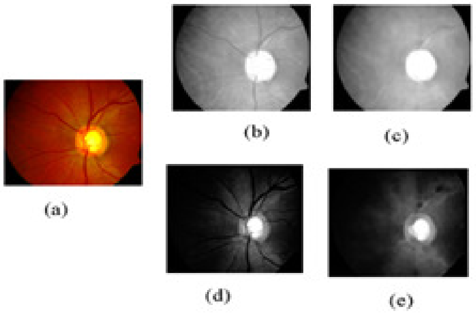


Figure 2. The preprocessing steps (a) Input Image (b) Red channel (c) Vessel free Image (d) Green channel (e) Vessel free Image

4.2. Feature Extraction

Medical image segmentation is a difficult task due to the complexity of segmentation. Because of its simplicity and efficiency, threshold segmentation is wildly used in many fields. Assessment of OD and OC is important in discriminating between normal and pathological retinal images. The OD is a bright pattern of the fundus image. Recently, many studies on the use of fundus images in extracting OD and OC have been reported. Fuzzy C Means Clustering with thresholding is used in this work for the extraction of OD and OC.

4.3. Fuzzy C Means Clustering with Thresholding

The proposed method is a combination of fuzzy algorithm, C Means clustering and thresholding. Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Different types of similarity measures may be used to identify classes depending on the data and the application, where the similarity measure controls the formation of the clusters. In the following new method intensity value is used as the similarity measure. Thresholding is one of the most powerful techniques for image segmentation, in which the pixels are partitioned depending on their intensity value.

4.3.1. Fuzzy C-Means Clustering Algorithm

Fuzzy C-Means (FCM) Clustering is a clustering technique and it employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. It is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function. Corresponding to each cluster center, this algorithm works by assigning membership to each data point on the basis of the difference between the cluster center and the data point. The more the data is near to the cluster center, the more is its membership towards the particular cluster center. It is obvious that the summation of membership of each data point should be equal to one.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,...,n \qquad (3)$$

$$d_{ik} = \quad [\sum^{m}_{j=1} [x_{kj} - v_j]^2]^{1/2} , \qquad (4)$$

where x_{kj} is data element, d_{ik} is the distance matrix and v_{ij} is the element of the cluster center vector.

The dissimilarity function which is used in FCM is given Equation (5)

$$J(U,c_1,c_2,...,c_c)=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}{}^{m}d_{ij}{}^{2} \qquad (5)$$

uij is between 0 and 1;

ci is the centroid of cluster i;

dij is the Euclidian distance between ith centroid(ci) and jth data point;

m ϵ [1,∞] is a weighting exponent.

To reach a minimum of dissimilarity function there are two conditions. These are given in Equation (6) and Equation (7)

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}{}^{m}x_j}{\sum_{j=1}^{n}u_{ij}{}^{m}} \qquad (6)$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \qquad (7)$$

This algorithm determines the following steps [4].

Step1. Randomly initialize the membership matrix (U) that has constraints in Equation 7.

Step2. Calculate centroids (C_i) by using Equation(6).

Step3. Compute dissimilarity between centroids and data points using equation (5). Stop if its improvement over previous iteration is below a threshold.

Step4. Compute a new U using Equation(7). Go to Step 2 [16][17].

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the apt location within a data set. To accommodate the introduction of fuzzy partitioning, the membership matrix (U) is randomly initialized according to Equation (7).

The Fuzzy Logic Toolbox command line function *fcm* is used for generating clusters, and in this paper three clusters are generated from the vessel free enhanced image. The *fcm* function iteratively moves the cluster centers to the right location within the data set. The outputs are 3 cluster centers C1, C2 and C3 and membership function matrix M with membership-grades, which is the intensity value of each pixel.

4.4. Thresholding

Thresholding is the operation of converting a multilevel image into a binary image i.e., it assigns the value of 0 (background) or 1 (objects or foreground) to each pixel of an image based on a comparison with some threshold value T (intensity or colour value) [13][14][15]. By applying the threshold T on an image, the image is converted to a binary image. The following formula (8) [13] is used for the binary image extraction.

$$I_T(x, y) = \begin{cases} 1, \text{if } \; I(x, y) > T \\ 0, \text{if } \; I(x, y) <= T, \end{cases} \qquad (8)$$

where *I* is the input image, T is the threshold and I_T is the binary image after thresholding.

4.5. Extraction of Optic Disc

The main feature of the OD is that it is having the highest intensity. Therefore the highest intensity is used as the threshold for the OD extraction. The threshold T is computed using the following method. From the generated clusters, first the cluster with maximum membership grade is found, and the corresponding grades are assigned with the same identification label. From the smoothed image, pixels with this gray level value are accessed, the average of the maximum and minimum intensity values are computed to obtain the threshold value T_1.

$$\text{i.e.,} \quad T_1 = \frac{1}{2}[\text{Max (data (value))} + \text{Min (data (value))}] \qquad (9)$$

In the above equation, data represents the data points of the smoothed red channel image and label represents the cluster value with the highest membership grade. By applying the threshold T_1 on the smoothed image I_S the image is converted to a binary image. The formula (9) is used for the binary image extraction.

Since the OD is of circular shape, the OD region selection process needs to be made specific to the circular region. So the largest connected component *Ri* whose shape is approximately circular is selected using the compactness measure

$$C\,(Ri) = \frac{P(Ri)}{4\pi A(Ri)} \qquad (10),$$

where, P(Ri) is the perimeter of the region *Ri* and A(Ri) is the area of the region *Ri*. The binary image with the compactness smaller than the pre-specified value, (5 in the present study) is considered as the optic disc approximation. Thus using the condition C < 5, extraction of round objects is done, eliminating those objects that do not meet the criteria. In some cases the extracted image contains small unwanted objects. The erosion operation is used to remove these objects.
The mean of the rows and columns form the centroid (Y1, X1) of the OD.

$$Y1 = \frac{1}{m}\sum_{i=1}^{m} \text{row1i} \qquad (11)$$

$$X1 = \frac{1}{n}\sum_{i=1}^{n} \text{col1i}\,, \qquad (12)$$

where *m* is the number of rows and *n* is the number of columns.

From the above coordinates of the optic disc the minimum coordinates (ymin1, xmin1) is calculated. The distance between the centroid and (ymin1, xmin1) represents the radius of the disc.

$$R_{OD} = Y1 - ymin1 \quad , \qquad (13)$$

where R_{OD} is the radius of the optic disc.

4.6. Optic Disc Segmentation

OD segmentation obtains a circular boundary approximation within a retinal image. A circle is plotted using the centroid (Y1, X1) and radius R_{OD}, gives segmented OD on the colour fundus eye image. Figure 3 shows the segmented OD [18].

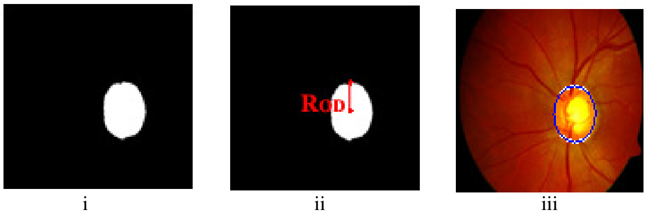


Figure 3 i. Extracted optic disc ii. Centre and radius of extracted disc iii. Optic disc segmentation

4.7. Extraction of Optic Cup

The above mentioned FCM clustering with thresholding is applied on the smoothed green channel for the extraction of OC.

The following algorithm includes four steps [4].

Step1. Randomly initialize the membership matrix (U) that has constraints in Equation (6).

Step2. Calculate centroids (Ci) by using Equation (7).

Step3. Compute dissimilarity between centroids and data points using equation (5). Stop if its improvement over previous iteration is below a threshold.

Step4. Compute a new U using Equation (6). Go to Step 2[16][17].

The threshold values T_2 is calculated using the following equation.

$$T_2 = \frac{1}{2}[\text{Max (data (value))} + \text{Min (data (value))}], \qquad (14)$$

where data represents the data points of the vessel free green channel and label represents the cluster value with the highest membership grade.

Since the OC is the brightest portion in the green channel, thresholding with threshold T_2 in im2bw function helps to extract OC. This function returns the binary image forming the object OC.

The average of the rows and columns forms the centroid (Y2, X2) of the OC.

$$\mathrm{Y2} = \frac{1}{\mathrm{m2}} \sum_{\mathrm{i=1}}^{\mathrm{m2}} \mathrm{row2i} \qquad (15)$$

$$\mathrm{X2} = \frac{1}{\mathrm{n2}} \sum_{\mathrm{i=1}}^{\mathrm{n2}} \mathrm{col2i}\ , \qquad (16)$$

where *m2* is the number of rows and *n2* is the number of columns.

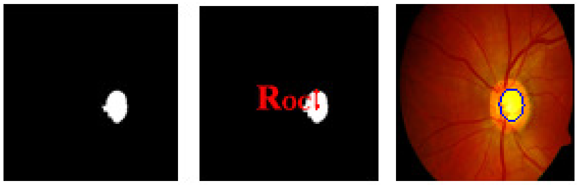
From the above coordinates of the optic cup the minimum coordinates (ymin2, xmin2) is calculated. The Euclidian distance between the centroid and (ymin1, xmin1) returns the radius of the cup.

$$\mathrm{R_{OC}} = \mathrm{Y2} - \mathrm{ymin2}, \qquad (17)$$

where $\mathrm{R_{OC}}$ is the radius of the cup.

4.8. Optic Cup Segmentation

OC within OD usually appears in circular shape. Therefore the OC segmentation is a circular boundary approximation. Using the centroid (Y2, X2) and radius $\mathrm{R_{OC}}$ a circle is drawn onto the current axes of the input image which would give the segmented OC on the colour fundus eye image.



i ii iii

Figure 4 i. Extracted Optic cup ii. Localization of centre and radius of optic cup iii. Optic cup segmentation

4.9. Computation of CDR

The manual method uses the ratio of the vertical diameter of OC and OD for the computation of CDR. From the segmented OD the minimum row coordinate $y_{min1}$and maximum row coordinate y_{max1} are calculated. The Euclidian distance between these coordinates is the vertical diameter of the OD, OD*vdiam* .

$$\mathrm{OD}vdiam \ = ymax1 - ymin1 \qquad (18)$$

Similarly from the segmented OC the minimum row coordinate $y_{min2}$and maximum row coordinate y_{max2} are calculated. The Euclidian distance between these coordinates is the vertical diameter of the OC, OC*vdiam*.

$$\mathrm{OC}vdiam \ = ymax2 - ymin2 \qquad (19)$$

The CDR is calculated using the following formula

$$\mathrm{CDR} = \mathrm{OC}vdiam \ / \ \mathrm{OD}vdiam \qquad (20)$$

The following figure shows the OD vertical diameter OD*vdiam* and OC vertical diameter OC*vdiam* of the input image.

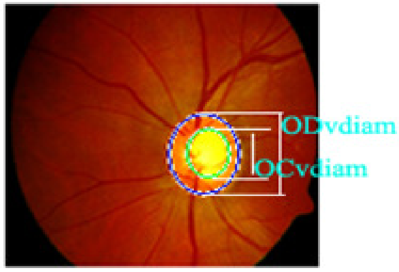


Figure5 Optic Disc vertical diameter ODvdiam and Optic Cup vertical diameter OCvdiam

5. RESULTS AND DISCUSSION

The automatic detection and evaluation of OD and OC is required for automatic diagnosis using retinal images. This study thus brings to light simple but efficient methods for the extraction and segmentation of OD and OC in retinal images. The CDR values are also automatically calculated. The new method is evaluated on the basis of the ground truth data, where, vertical CDR values are obtained from an expert ophthalmologist. Four hundred and fifty four color retinal images, including thirty normal and four hundred and thirty three pathological images, are used in this test. The performance evaluation is done by making use of the scatter plot analysis.

5.1 Image Data Sets

4.1.1 The DIARETDB0 and DIARETDB1 Databases

The DIARETDB0 and DIARETDB1 Database images were captured using an FOV of 50° and the size of each image is 1500 x 1152 x 3. Out of the 130 images of the DIARETDB0 database, 20 have normal architecture and 110 have various types of pathology. Out of the 89 images of the DIARETDB1 database, 5 have normal architecture and 84 have various types of pathology.

5.1.2 DRION Database

It has 110 retinal images with each image having the resolution of 600 x 400 pixels and the optic disc annotated by two experts with 36 landmarks. The mean age of the patients was 53.0 years (standard Deviation 13.05), with 46.2% male and 53.8% female and all of them were Caucasian ethnicity 23.1% patients had chronic simple glaucoma and 76.9% eye hypertension. The images were acquired with a colour analogical fundus camera, approximately centered on the ONH and they were stored in slide format. In order to have the images in digital format, they were digitized using a HP-PhotoSmart-S20 high-resolution scanner, RGB format, resolution 600x400 and 8 bits/pixel.

5.1.3 Images from the ophthalmologist

125 images from Giridhar Eye Institute, Kochi was also used in this paper. All the images were obtained using Carlzeiss fundus camera. In total 5 are normal images and remaining 120 are diseased and the size of each image is 576 x 768 x 3.

5.2 Implementation

The new algorithm was applied on 454 images obtained from the above mentioned databases and ophthalmologists. Seven of the input images from each data set, along with their OD and OD segmentation on the input image, is shown in fig.6 (a) and fig.6 (b) respectively.

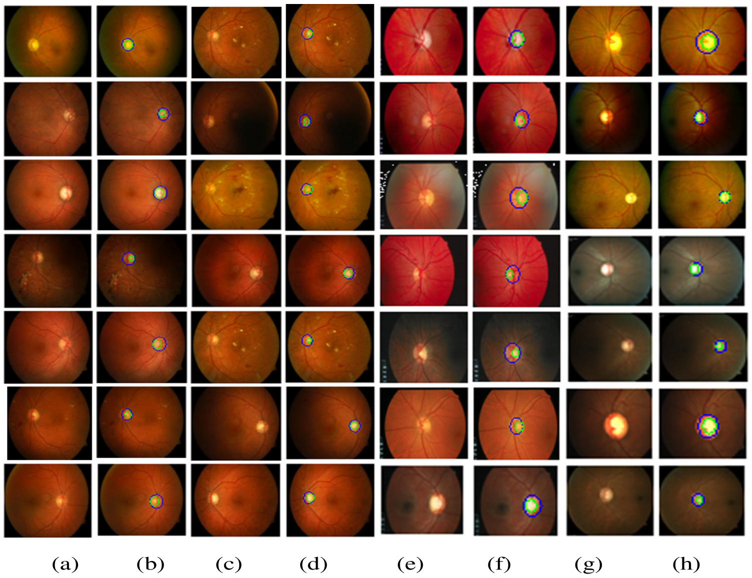


Figure 6 Input Image and OD & OC segmentation using new method on images from (a) and(b) DIARATDB0 (e) and(f) DRION (c) and(d)

5.3 Performance Evaluation

The performance evaluation is done using the following parameters.

5.3.1 Success rate

The decision for successful segmentation or failed segmentation is based on human eye observation. Table 1 shows the success rate of OD and OC segmentation using 454 images.

Table1. Success Rate Table

Database	Normal	Pathological	Total	Success Rate (%)
Drion	0	110	110	94.5
Diaretdb0	20	110	130	95.4
Diaretdb1	5	84	89	93.3
Ophthalmologist	5	129	125	97.3
Total	30	433	454	94.26 (Average)

5.3.2 Accuracy

The accuracy of the technique was evaluated quantitatively by comparing the obtained vertical CDR values with ophthalmologists' ground-truth vertical CDR values. Fifteen examples of detailed results of performance measurement using FCM clustering combined with thresholding are displayed in Table II using fifteen test images of DRION database and fifteen test images from the ophthalmologist.

Table II CDR Comparison Table shows the comparison of clinical CDR values with CDR values obtained using the new method.

Images	**Clinical CDR (1)**	**Obtained CDR** (2)	**Difference (1) –(2)**	**Clinical CDR of DRION Database (3)**	**Obtained CDR (4)**	**Difference (3)-(4)**
Image 1	0.5000	0.6082	0.1082	0.3333	0.4000	0.0667
Image 2	0.5714	0.6231	0.0517	0.5734	0.5310	0.0424
Image 3	0.6666	0.5505	0.1161	0.6666	0.6080	0.0586
Image 4	0.8517	0.7871	0.0646	0.7000	0.7543	0.0543
Image 5	0.7142	0.6412	0.073	0.6578	0.5816	0.0762
Image 6	0.4864	0.4173	0.0691	0.5625	0.5045	0.0580
Image 7	0.7060	0.6275	0.0785	0.4062	0.4995	0.0933
Image 8	0.9000	0.8367	0.0633	0.5000	0.6020	0.1020
Image 9	0.6801	0.7287	0.0486	0.6097	0.6696	0.0626
Image 10	0.9026	0.8206	0.0820	0.6410	0.6610	0.0200
Image 11	0.4631	0.4012	0.0619	0.5162	0.4011	0.1151
Image 12	0.6267	0.5151	0.1116	0.5010	0.5362	0.0352
Image 13	0.5147	0.4736	0.0411	0.6097	0.5215	0.0882
Image 14	0.7318	0.7180	0.0138	0.7408	0.6943	0.0465
Image 15	0.4265	0.3162	0.1103	0.6981	0.6208	0.0773
Mean Difference			0.07292	Mean Difference		0.06643

From the table it is shown that the mean differences between the clinical CDR values and obtained CDR values are very low.

5.3.3 Scatter plot Analysis

The statistical analysis is done using the scatter plot diagram. The clinical CDR vales and the obtained CDR values of the above mentioned data set are analysed using the scatter plot analysis. From the diagram it is shown that there exists highly positive linear relationship between both CDR values. Figure7 depicts the comparison between the clinical CDR and obtained CDR values.

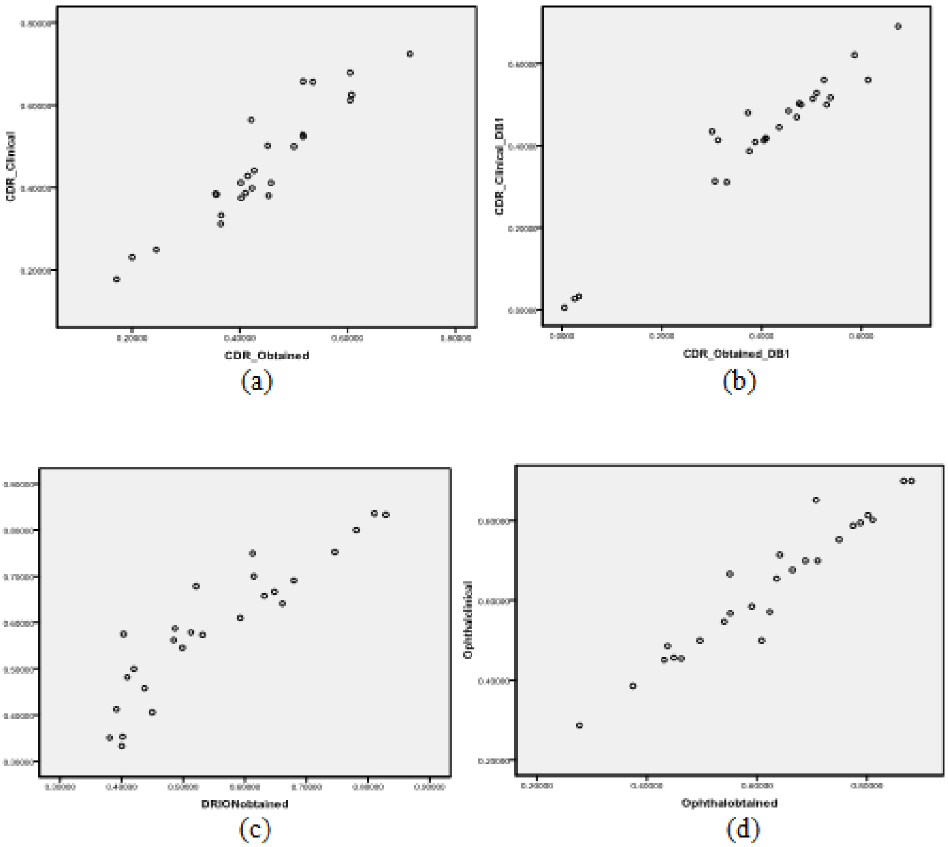


Figure7. Scatter Plot Diagram (a) Diaretdb0 (b) Diaretdb1 (c) Drion (d) Ophthalmologist

6. CONCLUSION

This paper presents a new fuzzy based approach for OD and OC extraction and segmentation together with CDR computation. The results presented in this paper show that the new methodology offers a reliable and robust solution for CDR computation. The scatter plot analysis results a high positive correlation between the clinical CDR and the obtained CDR. This automated method is very useful for the automatic screening of retinal images. However the present method has the following limitations. It is assumed that the OD and OC are brighter than the surrounding pixels and therefore cannot handle retinal images with a relatively dark OD. Hence advanced extraction methods are required for future studies and research.

REFERENCES

[1] Juan Xu, Hiroshi Ishikawa, Gadi Wollstein, Richard A. Bilonick, Kyung R. Sung,Larry Kagemann, Kelly A. Townsend, and Joel S. Schuman, “Automated Assessment of the Optic Nerve Head on Stereo Disc Photographs”, Invest Ophthalmol Vis Sci. Jun 2008; 49(6), pp. 2512–2517, 2008.

[2] Thitiporn Chanwimaluang and Guoliang Fan, “An efficient algorithm for extraction of anatomical structures in retinal images”, Proc. of International Conference on Image Processing, Vol. 1, pp. 1093–1096,2003

[3] Chisako Muramatsu, Toshiaki Nakagawab, Akira Sawadac, Yuji Hatanakad, Takeshi Haraa, Tetsuya Yamamotoc, Hiroshi Fujitaa “Determination of cup and disc ratio of optical nerve head for diagnosis

of glaucoma on stereo retinal fundus image pairs", Medical Imaging 2009: Computer-Aided Diagnosis, Proc. of SPIE ,Vol. 7260, pp.72603L-1 - 72603L-8, 2009

[4] Huiqi Li, Opas Chutatape "Boundary detection of optic disc by a modified ASM method", The Journal of the Pattern Recognition Society, Vol. 36, pp. 2093-2104,2003

[5] Huajun Ying, Ming Zhang and Jyh-Charn Liu, "Fractal-Based Automatic Localization and Segmentation of Optic Disc in Retinal Images", 28th Annual International Conference of the IEEE Engineering In Medicine and Biology Society (EMBS), 2007

[6] S. Sekhar, "Automated Localization Of Optic Disk And Images", Proc. 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008.

[7] Hussain, A.R., "Optic nerve head segmentation using genetic active Contours", International Conference on Computer and Communication Engineering, 2008,ICCCE 2008, 13-15 May 2008, pp.783 – 787, Univ. Teknikal Malaysia, Melaka, May 2008.

[8] Zhuo Zhang ,Jiang Liu , Cherian, N.S. ,Ying Sun, Joo Hwee Lim , Wing Kee Wong , Ngan Meng Tan ,Shijian Lu , Huiqi Li , Tien Ying Wong, "Convex hull based neuro-retinal optic cup ellipse optimization in glaucoma diagnosis", Engineering in Medicine and Biology Society, 2009. EMBC2009. Annual International Conference of the IEEE, Inst. for Infocomm Res., A*STAR, Singapore, 3-6 Sept. 2009, pp. 1441 – 1444, 2009.

[9] Wong, D.W.K. ; Jiang Liu ; Joo Hwee Lim ; Ngan Meng Tan ; Zhuo Zhang ; Huiqi Li ; Shijian Lu ; Tien Yin Wong , 'Method of detecting kink-bearing vessels in a retinal fundus image(CDR)',The 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1690 – 1694, June 2010

[10] Joshi, G.D., Sivaswamy, J., Karan, K., Prashanth, R., Krishnadas, S.R., "Vessel Bend-Based Cup Segmentation in Retinal Images", 20th International Conference on Pattern Recognition (ICPR), 2010, CVIT, IIIT Hyderabad, Hyderabad, India, 23-26 Aug. 2010, pp. 2536 – 2539, 2010.

[11] Shijian Lu and Joo Hwee Lim, "Automatic optic disc detection through background estimation" Proceedings of 2010 IEEE 17th International Conference on Image Processing, Hong Kong, September 26-29, 2010.

[12] Angel Suero, Diego Marin, Manuel E. Gegundez-Arias, and Jose M. Bravo, "Locating the Optic Disc in Retinal Images Using Morphological Techniques", IWBBIO 2013 Proceedings, Granada, 18-20 March, 2013, pp.593-600, 2013.

[13] Rafael C Gonzalez, Richard E Woods, Steven L Eddins, Digital Image Processing, Prentice Hall Publications,2008.

[14] N. M. Salem and A. K. Nandi, "Novel and adaptive contribution of the red channel in pre-processing of colour fundus images," in Journal of the Franklin Institute, 2007, p. 243256, 2007.

[15] Rafael C Gonzalez, Richard E Woods, Steven L Eddins., Digital Image Processing Using Matlab, Prentice Hall Publications,2008.

[16] HeikoTimm, Christian Borgelt, and Rudolf KruseFuzzy, "Cluster Analysis with Cluster Repulsion", CiteSeerx.

[17] Yinghua Lu, Tinghuai Ma, Changhong Yin, Xiaoyu Xie, Wei Tian,ShuiMing Zhong, "Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data", International Journal of Database Theory and Application, Vol.6, No.6, pp.1-18, 2013

[18] Thresiamma Devasia, Poulose Jacob, Tessamma Thomas, "Automatic Optic Disc Boundary Extraction from Color Fundus Images", International Journal of Advanced Computer Science and Applications (IJACSA) , Vol. 5, No. 7, 2014,pp.17-24

[19] Yong Yang , Shuying Huang, " Image segmentation by fuzzy c-means Clustering algorithm with a novel Penalty term", Computing and Informatics, Vol. 26, pp. 17–31, 2007.

AUTHORS

Thresiamma Devasia was graduated with Bachelor of Mathematics (BSc.Maths) from Mahatma University, Kerala, India in 1991, and finished her Master of Computer Applications (MCA) and M.Phil Computer Science from Alagappa University Tamilnadu, India in 1995 and 2010, respectively.

Currently, she is the Head and Associate professor, Department of Computer Science at Assumption College Changanacherry, Kerala, India and working toward her Ph.D. at Cochin University of Science And Technology on glaucoma detection using image processing.

She completed UGC sponsored minor research project based on image processing. She was a member of IEEE. Her interest areas include image processing and medical imaging.

Dr. K.Poulose Jacob, Professor of Computer Science at Cochin University of Science and Technology since 1994, is currently Pro Vice Chancellor of Cochin University of Science & Technology. He has presented research papers in several International Conferences in Europe, USA, UK, Australia and other countries. He has served as a Member of the Standing Committee of the UGC on Computer Education & Development. He is the Zonal Coordinator of the DOEACC Society under the Ministry of Information Technology, Government of India. He serves as a member of the AICTE expert panel for accreditation and approval. He has been a member of several academic bodies of different Universities and Institutes. He is on the editorial board of two international journals in Computer Science. Dr. K.Poulose Jacob is a Professional member of the ACM (Association for Computing Machinery) and a Life Member of the Computer Society of India.

Dr.Tessamma Thomas received her M.Tech. and Ph.D from Cochin University of Science and Technology, Cochin-22, India. At present she is working as Professor in the Department of Electronics, Cochin University of Science and Technology. She has to her credit more than 100 research papers, in various research fields, published in International and National journals and conferences. Her areas of interest include digital signal / image processing, bio medical image processing, super resolution, content based image retrieval, genomic signal processing, etc.

COMPARATIVE STUDY OF DIMENSIONALITY REDUCTION TECHNIQUES USING PCA AND LDA FOR CONTENT BASED IMAGE RETRIEVAL

Shereena V. B.[1]and Julie M. David[2]

[1,2] Asst. Professor, Dept. of Computer Applications, MES College,
Marampally, Aluva, Cochin, India
[1]shereenavb@gmail.com
[2]julieeldhosem@yahoo.com

ABSTRACT

The aim of this paper is to present a comparative study of two linear dimension reduction methods namely PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). The main idea of PCA is to transform the high dimensional input space onto the feature space where the maximal variance is displayed. The feature selection in traditional LDA is obtained by maximizing the difference between classes and minimizing the distance within classes. PCA finds the axes with maximum variance for the whole data set where LDA tries to find the axes for best class seperability. The proposed method is experimented over a general image database using Matlab. The performance of these systems has been evaluated by Precision and Recall measures. Experimental results show that PCA based dimension reduction method gives the better performance in terms of higher precision and recall values with lesser computational complexity than the LDA based method.

KEYWORDS

Color histogram, Feature Extraction, Euclidean distance, Principal Component Analysis, Linear Discriminant Analysis, Eigen Values, Eigen Vectors.

1. INTRODUCTION

As we know human beings are predominantly visual creatures. The visualisation of the images which we see, in real or imaginary, make sense of the world around us to identify and differentiate the things which we see at a quick glance. We are bestowed with very precise visual skills to identify an image by size and also by differentiating the colors. We can process a large amount of visual information very quickly.

An image processing task consists of acquiring the image, pre-processing, segmentation, representation and description and finally recognition and interpretation. There are four types of digital images, binary, grey scale, true color or RGB and indexed [1]. Binary representation images include text, fingerprints or architectural plans where each pixel is black or white. Grey

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 41–55, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50905

scale images consist of X-rays, images of printed works etc where each pixel is a shade of grey, normally from 0 to 255. True color or RGB images are the color images where each pixel is described by the amount of red, green and blue in it. Finally there are indexed images where the image has an associated color map which is a list of all the colors used in that image. Each pixel has a value which does not give its color, but an index to the color in the map.

There has been a tremendous growth in the digital information over years. This trend has motivated research in image databases, which were nearly ignored by traditional computer systems due to the enormous amount of data necessary to represent images and the difficulty of automatically analyzing images. Currently, storage is less of an issue since huge storage capacity is available at low cost. Large image databases are used in many application areas such as satellite imaging, and biometric databases, Crime prevention, military, Intellectual property, Architectural and engineering design, Fashion and interior design, Journalism and advertising, Medical diagnosis, Geographical information and remote sensing systems, Cultural heritage, Education and training, Home entertainment, Web searching, where it is important to maintain a high degree of precision [2]. Thus an important issue was the fast image retrieval from large databases. This trend led to the development of research area known as Content Based Image Retrieval. CBIR systems retrieves features from the raw images themselves and calculate an association measure between the query image and database images based on these features. We need to develop an efficient system for retrieving images since speed and precision are important.

CBIR consists of different stages such as Image acquisition, image Pre-Processing, Feature Extraction, Similarity Matching and obtain the resultant images. Image Acquisition is the process of acquiring a digital image database which consists of n number of images. The Pre-processing stage involves filtering, normalization, segmentation, and object identification. The output of this stage is a set of significant regions and objects. In the Feature extraction stage, visual information such as color and texture is extracted from the images and saves them as feature vectors in a feature vector database. One of the major problems with Content Based image retrieval system is the large number of features extracted which requires large amount of memory and computation power. To overcome this problem we have to construct a combination of features which best describe the data with sufficient accuracy. So in this stage, we use dimension reduction algorithms which extract only essential features from the feature vector database and store them as reduced feature vector database. Thus the output of feature extraction stage is a reduced set of features which best describes the image. In the Similarity matching stage, the reduced feature vectors of query image calculated is matched with the feature vectors of reduced feature vector database using any of the Distance methods available such as Euclidean distance, City Block Distance, Canberra Distance [3].

The most popular among the Dimensionality Reduction Algorithms are Principal Component Analysis and Linear Discriminant Analysis. Principal Component Analysis defines new attributes (principal components or PCs) as mutually-orthogonal linear combinations of the original attributes. For many image datasets, it is sufficient to consider only the first few PCs, thus reducing the dimension. Linear Discriminant Analysis [4] easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. In this paper, we compare the above two dimensionality reduction techniques by implementing the algorithms on a given image data set. Experimental results on database images shows that the feature set can be considerably reduced without significant degradation in performance.

The rest of this paper is organized as follows. Section 2 deals with Literature Review. In Section 3, we explain Proposed Methodology. Section 4 consists of Comparative study of PCA and LDA, Conclusions are given in Section 5.

2. LITERATURE REVIEW

H.H. Pavan Kumar Bhuravarjula and VNS Vijayakumar proposed in their paper "A novel content based image retrieval using variance color moment" that color moments gives average high precision and recall [2]. In the paper of Manimala Singha and K. Hemachandran [5], they presented a novel approach for Content Based Image Retrieval by combining the color and texture features called Wavelet-Based Color Histogram Image Retrieval (WBCHIR). The experimental result shows that the proposed method outperforms the other retrieval methods in terms of Average Precision. Pranali Prakash Lokhande , P. A. Tijare [6] concluded in their paper " Feature Extraction Approach for Content Based Image Retrieval "that the combination of the color and texture features of an image in conjunction with the shape features will provide a robust feature set for image retrieval. S. Mangijao Singh and K. Hemachandran [7] in their paper "Content-Based Image Retrieval using Color Moment and Gabor Texture Feature" proposed an efficient image retrieval method based on color moments and Gabor texture features. To improve the discriminating power of color indexing techniques, they encoded a minimal amount of spatial information in the index. Mohd. Danish, Ritika Rawat, Ratika Sharma [3] in their paper "A Survey: Content Based Image Retrieval Based On Color, Texture, Shape and Neuro Fuzzy" provides an overview of the functionality of content based image retrieval systems. Most systems use color and texture features, and some systems use shape features.

A. Ramesh Kumar and D. Saravanan in their paper "Content Based Image Retrieval Using Color Histogram" [8], CBIR using color histograms technique is proposed with help of principal component analysis technique to improve the image retrieval performance. Swati V. Sakhare and Vrushali G. Nasre, [9] in their paper "Design of Feature Extraction in Content Based Image Retrieval (CBIR) using Color and Texture" designed an application which performs a simple color-based search in an image database for an input query image, using color, texture and shape to give the images which are similar to the input image as the output. The number of search results may vary depending on the number of similar images in the database. In the paper "A Proposed Method for Image Retrieval using Histogram values and Texture Descriptor Analysis" [10], Wasim Khan, Shiv Kumar. Neetesh Gupta and Nilofar Khan proposed a method for image retrieval using histogram values and texture descriptor analysis of image. When a query image is submitted, its color and texture value is compared with the color and texture value of different images stored in database. The images having closest value compared to query image are retrieved from database are displayed on GUI as result.

S. Meenachi Sundaresan and Dr. K.G. Srinivasagan [11] proposed in their paper "Design of Image Retrieval Efficacy System Based on CBIR" that the performance of a retrieval system can be measured in terms of its recall (or sensitivity) and precision (or specificity). Recall measures the ability of the system to retrieve all models that are relevant, while precision measures the ability of the system to retrieve only models that are relevant. In the paper " An Enhancement on Content-Based Image Retrieval using Color and Texture Features", [12] Tamer Mehyar, Jalal Omer Atoum proposed an enhancement on the use of color and texture visual features in Content-Based Image Retrieval (CBIR) by adding a new color feature called Average Color Dominance which tries to enhance color description using the dominant colors of an image.

In the paper “Implementation of Principal Component Analysis with Fuzzy Annotation for CAD Jewellery Images”, Pinderjeet Kaur [13] proposed that Principal Component Analysis (PCA) can be used for dimension reduction to reduce the computation cost for the system of Content Based Image Retrieval (CBIR). Arunasakthi. K, KamatchiPriya. L [14] stated in their paper “A Review On Linear And Non-Linear Dimensionality Reduction Techniques” that Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are regarded as the most fundamental and powerful tools of dimensionality reduction for extracting effective features of high-dimensional vectors in input data. According to Julie M. David and Kannan Balakrishanan, principal components are new set of variables which are generated by the application of dimensionality reduction method [15]. The basic procedures behind PCA are (i) the inputs data are normalized, so that each attribute falls within the same range. This helps ensure that attributes with large domains will not dominate attributes with smaller domains, (ii) PCA computes k orthonormal vectors that provides a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components and (iii) The principal components are sorted in order of decreasing strength. [16]

Kresimir Delac, Mislav Grgic and Sonja Grgic [17] in their paper “Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set” proposed that PCA finds a set of the most representative projection vectors such that the projected samples retain most information about original samples whereas LDA uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter. In the paper “CBIR Feature Vector Dimension Reduction with Eigenvectors of Covariance Matrix using Row, Column and Diagonal Mean Sequences” [18], Dr. H.B. Kekre, Sudeep D. Thepade and Akshay Maloo stated that PCA can be used to transform each original image from database into its corresponding eigen image.

In the paper “Linear Discriminant Analysis bit by bit” Sebastian Raschka [19] stated that PCA can be described as an unsupervised algorithm, since it ignores class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is supervised and computes the linear discriminants that will represent the axes that maximize the separation between multiple classes.

The main motivation of this work is to compare two dimension reduction techniques PCA and LDA to find out which of them selects the best features from the feature set to reduce the dimensions of the dataset with minimal loss of information. Principal Component Analysis (PCA) is a mathematical tool used to extract principal components of original image data. These principal components may also be referred as Eigen images. Linear Discriminant Analysis seeks to reduce dimensionality while preserving as much of the class discriminatory information as possible. In LDA, we compute eigenvectors from our dataset and collect them in scatter matrices.

3. PROPOSED METHODOLOGY

3.1 Prepare input data.

In this paper, a general image database consisting of 500 images is used for testing the comparative study of PCA and LDA. Principal Component Analysis defines new attributes as mutually-orthogonal linear combinations of the original attributes. Linear Discriminant Analysis computes the linear discriminants that will represent the axes that maximize the separation between multiple classes. In order to obtain better search results and to express more image information, we consider the dominant color and texture features combined. These low level features are extracted using color moments, color histogram, color auto correlogram and wavelet. The basis of color moments is that the distribution of color in an image can be considered as a probability distribution which can be characterized by various moments [20]. The color histogram for an image is constructed by quantizing the colors within the image and counting the number of pixels of each color. The color correlogram was proposed to characterize not only the color distributions of pixels, but also the spatial correlation of pairs of colors. Wavelet Analysis is a popular method for extracting texture from an image. The discreet wavelet transform (DWT) of a signal is calculated by passing it through a series of filters (high and low pass filters) and then down-sampled [21].

3.1.1 Color Moments

The first order color moment(Mean), Second order color moment(Standard deviation) and the third order color moment(Skewness) have been used for color feature extraction [20]. Since only 9 (three moments for each of the three color components R, G, B) numbers are used to represent the color content of each image, color moments are a very compact representation compared to other color features.

$$\mu_i = \frac{1}{N}\sum_{j=1}^{n} P_{ij}$$

$$\sigma_i = (\frac{1}{N}\sum_{j=1}^{n}((P_{ij} - \mu_i)^2)^{\frac{1}{2}}$$

$$S_i = (\frac{1}{N}\sum_{j=1}^{n}((P_{ij} - \mu_i)^3)^{\frac{1}{3}}$$

Where P_{ij} is the value of the i- the color channel of image pixel j and N is the number of pixels in the image.

When a query image is submitted for image retrieval, its color moments are extracted and added to feature set for matching the image with the images stored in the database. The following are the steps for extracting color moments from an image.

1. Extract the values of each plane R, G, B corresponding to the image.
2. Find the mean , standard deviation and skewness of each plane
3. Convert to column vector output of the moments.

The following table gives the color moments of 5 images where M corresponds to mean, Std corresponds to standard deviation, Skew corresponds to Skewness and R for Red, G for Green and B for Blue plane respectively.

Table 1. Color Moments Table

M(R)	Std(R)	Skew(R)	M(G)	Std(G)	Skew(G)	M(B)	Std(B)	Skew(B)
0.4372	0.3659	0.2132	0.1925	0.1821	0.3083	0.0014	0.0013	0.0009
0.4385	0.3409	0.2389	0.2018	0.1928	0.3131	0.0011	0.0015	0.0010
0.4351	0.3572	0.2582	0.2521	0.2349	0.3069	0.0019	0.0033	0.0030
0.5061	0.4364	0.2362	0.2355	0.2283	0.4019	0.0008	0.0015	0.0011
0.3765	0.4012	0.2818	0.2850	0.2844	0.3319	0.0054	0.0064	0.0046

3.1.2 Color Histogram

A histogram is a graph that represents all the colors and the level of their occurrence in an image irrespective of the type of the image [8]. This technique describes the proportion of pixels of each color in an image. It has been used as one of the feature extraction attributes with the advantage like robustness with respect to geometric changes of the objects in the image. The color histogram is obtained by quantizing image colors into discrete levels and then counting the number of times each discrete color occurs in the image. In a CBIR system, a query image is compared with the histograms of all the images in database [22].

A color histogram H for a given image is defined as a vector

$$H = \{ H[1], H[2], . . . H[i], . . . , H[N]\}$$

where i represent a color in the color histogram, H[i] is the number of pixels in color i in that image, and N is the number of bins in the color histogram, i.e., the number of colors in the adopted color model.

In order to compare images of different sizes, color histograms should be normalized. The normalized color histogram H ′ is defined as

$$H' = \{H'[0], H'[1], ...H'[i], .. H'[N]\}$$

where $H'[i] = \frac{H[i]}{XY}$, XY is the total number of pixels in an image.

From the query image submitted for image retrieval, its color histogram features are extracted and added to feature set for matching the image with database images. The following steps give a method to calculate color histogram.

1. Convert the image from RGB color space to HSV color space.
2. Define number of clusters for each HSV plane.
3. Find the maximum value of each plane.
4. Cluster each values after normalisation.
5. Add each color to any one of the appropriate cluster.
6. Find the probabilistic values and convert the values to the column vector.

3.1.3 Color autocorrelogram

A color correlogram is a table indexed by color pairs, where the k-th entry for (i, j) specifies the probability of finding a pixel of color j at a distance k from a pixel of color i in the image [20]. Let I represent the entire set of image pixels and $I_{c(i)}$ represent the set of pixels whose colors are c(i).Then, the color correlogram is defined as:

$$\gamma_{(i,j)}(k) = Pr_{p1\in c(I),p2\in I}\ \ [p2 \in I_{c(j)}|p1 - p2| = k]$$

Where i, j ∈ {1, 2, …, N}, k∈ {1, 2, …, d}, and | p1 – p2 | is the distance between pixels p1 and p2.

The color auto correlogram of the query image is extracted and added to feature vector for the extraction of similar database images. The following are the steps for extracting correlogram features from an image.

1. Reduce the number of colors in the RGB image.
2. Correlate each pixel with the neighbourhood pixels for getting the correlogram vector.

3.1.4 Texture

Like color, the texture is a powerful low-level feature for image search and retrieval applications. The texture measures try to retrieve the image or image parts characteristics with reference to the changes in certain directions and the scale of the images. This is most useful for images with homogeneous texture [3]. Wavelet analysis is an exciting new method for solving difficult problems in mathematics, physics, and engineering, with modern applications as wave propagation, data compression, signal processing, image processing, pattern recognition, computer graphics, the detection of aircraft and submarines and other medical image technology. A wavelet is a mathematical function used to divide a given function into different frequency components [21]. A wavelet transform is the representation of a function by wavelets, which represent scaled and translated copies of a finite length or fast-decaying oscillating waveform (known as the "mother wavelet"). The Wavelet transform of a function is the improved version of Fourier transform. Wavelet transforms have advantages over traditional Fourier transforms because local features can be described better with wavelets that have local extent. Some mother wavelet families implemented in Matlab are Daubechies, Symlet, Coiflet, Biortogonal and Reverse biorthogonal wavelets) and the fractional B-spline functions are used to compute different feature vectors. Orthogonal wavelets with FIR filters can be defined through a scaling filter. Predefined families of such wavelets include Haar, Daubechies, Symlets and Coiflets. In this paper, Coiflet wavelet function is used to extract texture features. The following steps give a method to calculate Texture of an image.

1. Convert the image to grayscale.
2. Find the 4 stage Coif wavelet coefficients.
3. Find the mean and standard deviation of the above coefficients and output to a column vector

The following table gives the 4 stage coiflet texture values of 5 images.

Table 2. Coiflet Texture values Table

5.9048	2.6054	0.1637	0.0979
5.5827	2.9509	0.2472	0.0167
6.2997	3.7201	0.2554	0.0405
7.1697	3.6297	0.2840	0.0528
6.5487	4.3823	0.0779	0.2445

The following table gives the first 10 features of 5 images in the database before applying dimension reduction algorithms.

Table3. Table of features before Dimension Reduction

1	2	3	4	5	6	7	8	9	10
0.4372	0.3659	0.2132	0.1925	0.1821	0.3083	0.0014	0.0013	0.0009	5.9048
0.4385	0.3409	0.2389	0.2018	0.1928	0.3131	0.0011	0.0015	0.0010	5.5827
0.4351	0.3572	0.2582	0.2521	0.2349	0.3069	0.0019	0.0033	0.0030	6.2997
0.5061	0.4364	0.2362	0.2355	0.2283	0.4019	0.0008	0.0015	0.0011	7.1697
0.3765	0.4012	0.2818	0.2850	0.2844	0.3319	0.0054	0.0064	0.0046	6.5487

3.2 Principal Component Analysis (PCA) Vs Linear Discriminant Analysis (LDA)

Principal Component Analysis is a technique which uses sophisticated underlying mathematical principles to transform a number of possibly correlated variables into a smaller number of variables called principal components [13]. It is one of the most important results from applied linear algebra. The advantage of PCA is finding the patterns in the data and compressing data by reducing the number of dimensions without loss of information. The mathematical concepts that are used for PCA are Standard Deviation, Variance, Co–variance and Eigenvectors [23]. The database images belonging to same category may differ in lighting conditions, noise etc., but are not completely random and in spite of their differences there may present some patterns. Such patterns could be referred as principal components. PCA is a mathematical tool used to extract principal components of original image data. These principal components may also be referred as Eigen images [18]. An important feature of PCA is that any original image from the image database can be reconstructed by combining the eigen images. The algorithm to calculate Principal Components is as follows.

1. Represent the image as one dimensional vector of size N x N.
   Suppose we have M vectors of size N (= rows of image × columns of image) representing a set of sampled images. Then the training set becomes: Γ1, Γ2, Γ3.....ΓM.

2. The Mean value of the pixels intensities in each image is calculated and subtracted from the corresponding image. The process is continued for all images in the database.

3. The covariance matrix which is of the order N^2 x N^2 is calculated as given by $C = AA^T$.

4. Find the Eigen values of the covariance matrix C by solving the equation $(C\lambda - I) = 0$ .To find the eigenvector X repeat the procedure where Xi indicates corresponding Eigen values.

5. The Eigen vectors are sorted according to the corresponding Eigen values in descending order.

6. Choose the First ‘K’ Eigen vectors and Eigen Values.

The following table gives the first 10 features of 5 images after applying the dimension reduction algorithm PCA and reducing the feature database.

Table 4. Table of features after Dimension Reduction

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
62.189	7.105e-15	-1.776e-15	1.154e-14	4.024e-15	-1.318e-15	-2.088e-15	-3.344e-15	9.992e-16	-1.089e-15
63.286	-2.442e-14	2.220e-15	3.996e-15	-1.554e-15	-3.885e-16	-2.396e-16	-1.707e-15	1.498e-15	9.436e-16
60.849	-3.497e-15	-1.065e-14	7.771e-16	2.220e-15	-2.657e-15	-5.568e-16	1.020e-15	-2.636e-15	3.677e-16
64.930	-1.776e-14	-5.329e-15	4.218e-15	1.110e-16	-1.845e-15	7.216e-16	-3.486e-16	-2.713e-15	-3.747e-15
51.539	8.882e-16	8.881e-15	-7.549e-15	-4.218e-15	5.343e-16	-2.331e-15	-2.307e-15	1.332e-15	1.637e-15

Linear Discriminant Analysis (LDA) [24] is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The feature selection in traditional LDA [14] is obtained by maximizing the difference between classes and minimizing the distance within classes. LDA finds the vectors in the underlying space that best discriminate among classes. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification [4].

1. Compute the *d*-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\mathbf{e_1, e_2, ..., e_d}$) and corresponding eigen values ($\boldsymbol{\lambda_1, \lambda_2, ..., \lambda_d}$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose **k** eigenvectors with the largest eigenvalues to form a $d \times k$-dimensional matrix **W** (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the mathematical equation: $\mathbf{y} = \mathbf{W}^{\mathbf{T}} \times \mathbf{x}$ (where **x** is a $d \times 1$-dimensional vector representing one sample, and **y** is the transformed $k \times 1$-dimensional sample in the new subspace).

3.3 Similarity Matching

If R' be the dimensionality reduced feature database and R" is the feature vector obtained from query image, then the retrieval system is based on a similarity measure defined between R' and R" [25]. In this paper, Euclidean distance is used to measure the similarity between the feature vectors of reduced query image and reduced database images. The formula for Euclidean distance [26] is given as

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{n}(Q_i - D_i)^2}$$

Where Q and D are feature vectors of the Query image and database image. After finding the Euclidean Distance, the distances are sorted and the top six images closer to the query image are retrieved.

3.4 Performance Evaluation

The performance of retrieval of the system can be measured in terms of its Recall and Precision. Recall measures the ability of the system to retrieve all the models that are relevant, while Precision measures the ability of the system to retrieve only the models that are relevant [20].

$$Precision = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ Number\ of\ images\ retrieved}$$

$$Recall = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ no\ of\ relevant\ images}$$

The number of relevant items retrieved is the number of the returned images that are similar to the query image in this case. The total number of images retrieved is the total number of images that are returned by the retrieval system. In precision and recall, crossover is the point on the graph where the both precision and recall curves meet. The higher the number of crossover points better will be the performance of the system.

4. COMPARATIVE STUDY OF PCA AND LDA

The proposed method has been implemented using Matlab 13 and tested on a general-purpose database containing 500 images, in JPG format of size 256X384 resized to 286x340. The database includes 500 color images categorized into five classes and each class includes 100 images as follows: African people, Beach, Building, Bus, Dinosaurs. The search is based on the similarity of feature vectors. We have followed the image retrieval technique, as described in the section 3 on different feature extraction schemes such as color and texture. This scheme calculated 110 features by means of histogram, moments, correlogram and Coif wavelet. Further, Principal Component Analysis technique and Linear Discriminant Analysis technique is used to extract the best features from the images. By means of PCA and LDA, the feature set is reduced to 75.Then the reduced query image is compared with the reduced database feature set using Euclidean Distance and the top 6 nearer images are displayed. The quality of the image retrieval, with different feature extraction schemes has been evaluated by randomly selecting query images, of each category, from test image database. Each query returns the top 6 images from database. To measure retrieval effectiveness for the image retrieval system, Precision and Recall values are used. The Precision Recall rates and plots for PCA, LDA and without dimension reduction methods are shown in figure1. The graphical user interface for the retrieval of images using dimension reduction with PCA and LDA are shown in the figure 2 and figure3 respectively. From the GUI, the database is to be selected first using Select Database button, i.e., the database of 500 images. Then the query image is selected from a set of test images using Select Query button. The query can be processed under 3 options- without dimension reduction, Dimension Reduction using PCA and Dimension Reduction using LDA. The images are retrieved based on the option selected and top 6 images are displayed in the Returned images frame. The Precision-Recall Plot gives the Precision and Recall rates of the selected option. The Performance Comparison button shows the Precision-Recall plots of all the three methods of the selected query image. From the Table 5 of Precision and Recall, it is found that the rates are higher for dimension reduction using Principal Component Analysis when compared to Linear Discriminant Analysis. This shows that PCA is a better dimension reduction tool when compared to LDA.

Table 5. Table of Precision Recall Rates

Class	**Without Dimension Reduction**		**LDA**		**PCA**	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Tribal	0.6110	0.5	0.6667	0.35	0.667	0.47
Beach	0.8330	0.46	0.5	0.25	0.778	0.37
Towers	0.6666	0.39	0.6667	0.37	0.667	0.37
Bus	0.722	0.4	0.333	0.27	0.667	0.31
Dinosaur	1	0.97	1	0.77	1	0.97

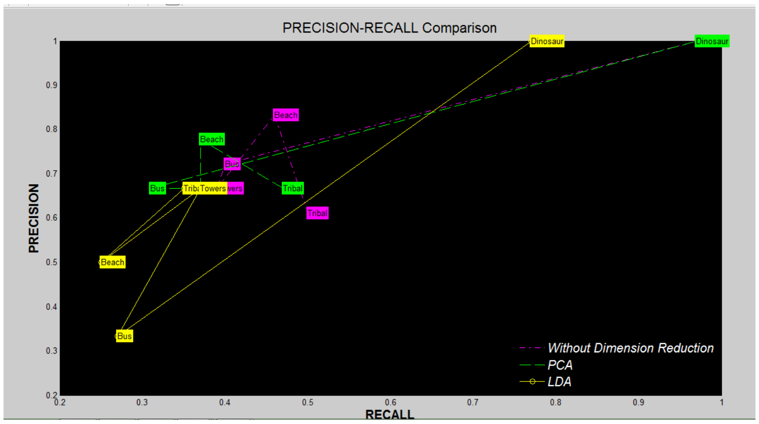
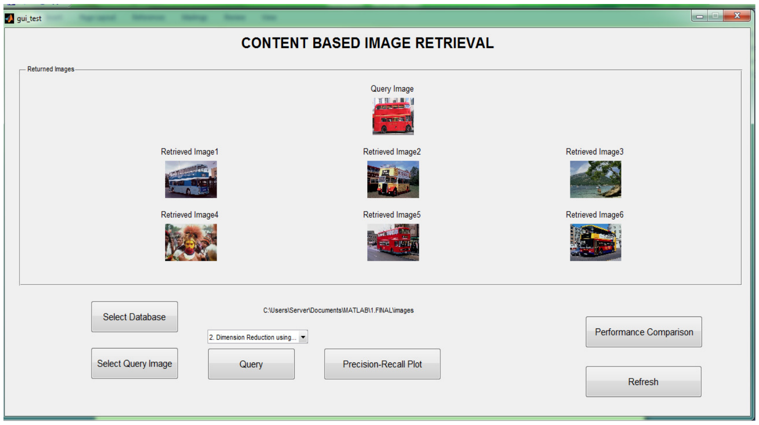


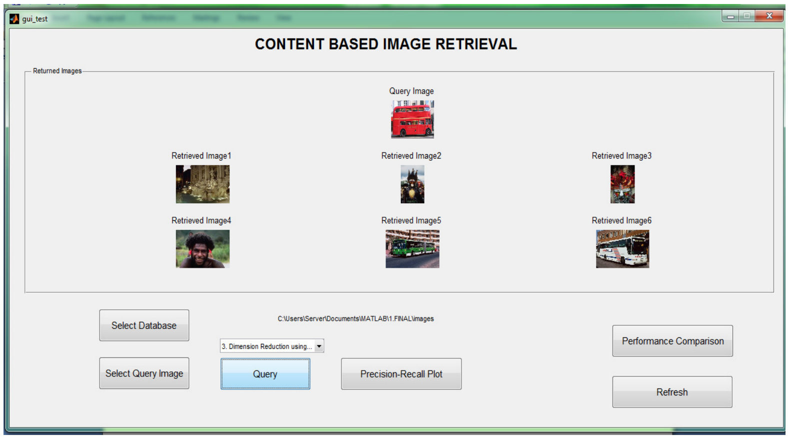
Figure1. Precision –Recall Plot



C:\Users\Server\Documents\MATLAB\1.FINAL\images

2. Dimension Reduction using...

Figure 2.Dimension Reduction using PCA



C:\Users\Server\Documents\MATLAB\1.FINAL\images

3. Dimension Reduction using...

Figure 3. Dimension Reduction using LDA

5. CONCLUSION

In this paper, we presented a comparative study of two dimension reduction methods namely Principal Component Analysis and Linear Discriminant Analysis. Dimensionality reduction methods aim at revealing meaningful structures and unexpected relationships in multivariate data. PCA projects correlated variables into a lower number of uncorrelated variables called principal components. By using only the first few principal components or eigen vectors, PCA makes it possible to reduce the number of significant dimensions of the data, while maintaining the maximum possible variance. The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. In LDA, we will compute eigenvectors from our dataset and collect them in scatter-matrices, the between-class scatter matrix and within-class scatter matrix. From the Precision and Recall rates of PCA and LDA calculated in the table in the above section, it can be found that the rates are high for all the cases of PCA when compared to LDA. Thus it is concluded that PCA tends to outperform LDA in almost all cases and hence PCA can be adopted as an effective tool for dimension reduction.

REFERENCES

[1] Rafael Gonzalez and Richard E. Woods, ‘Digital Image processing’, Addison Wesley, 2nd Edn., 2002.

[2] H.H. Pavan Kumar Bhuravarjula And V.N.S Vijaya Kumar(2012), ‘A Novel Content Based Image Retrieval Using Variance Color Moment’, Int. J. Of Computer And Electronics Research,Vol. 1, Issue 3, ISSN: 2278-5795, pp.93-99.

[3] Mohd. Danish, Ritika Rawat and Ratika Sharma(2013) , ‘A Survey: Content Based Image Retrieval Based On Color, Texture, Shape & Neuro Fuzzy’, Int. Journal Of Engineering Research And Applications ISSN : 2248-9622, Vol. 3, Issue 5, Sep-Oct 2013, pp.839-844 .

[4] S. Balakrishnama and A. Ganapathiraju, ‘Linear Discriminant Analysis - A Brief Tutorial’. Institute for Signal and Information Processing, Department of Electrical and Computer Engineering Mississippi State University.

[5] Manimala Singha and K. Hemachandran (2012) ‘Content based image retrieval using color and texture’, Signal & Image Processing : An Int. J. (SIPIJ), Vol.3, No.1, pp.39-57.

[6] Pranali Prakash Lokhande and P. A. Tijare(2012), ‘Feature Extraction Approach for Content Based Image Retrieval’ , Int.Journal of Advanced Research in Computer Science and Software Engineering , Vol. 2, Issue 2, ISSN: 2277 128X .

[7] S. Mangijao Singh , K. Hemachandran (2012) “Content-Based Image Retrieval using Color Moment and Gabor Texture Feature”, Int.J. of Computer Science Issues,Vol. 9, Issue 5, No 1, pp.299-309.

[8] A.Ramesh Kumar and D.Saravanan ,’ Content Based Image Retrieval Using Color Histogram’, Int. J. of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 242 – 245, ISSN:0975-9646.

[9] Swati V. Sakhare & Vrushali G. Nasre(2011), ‘Design of Feature Extraction in Content Based Image Retrieval (CBIR) using Color and Texture’, Int. J. of Computer Science & Informatics, Vol.-I, Issue-II, pp.57-61.

[10] Wasim Khan, Shiv Kumar. Neetesh Gupta and Nilofar Khan(2011), ‘A Proposed Method for Image Retrieval using Histogram values and Texture Descriptor Analysis’, Int. J.of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Vol.-I Issue-II pp.33-36.

[11] S.Meenachi Sundaresan and Dr. K.G.Srinivasagan(2013), ‘Design of Image Retrieval Efficacy System Based on CBIR’, Int. J. of Advanced Research in Computer Science and Software Engineering , Vol. 3, Issue 4, ISSN: 2277 128X, pp.48-53.

[12] Tamer Mehyar, Jalal Omer Atoum(2012), ‘An Enhancement on Content-Based Image Retrieval using Color and Texture Features’, Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 4,ISSN 2079-8407, pp488-496.

[13] Pinderjeet Kaur (2012), ‘Implementation of Principle Component Analysis with Fuzzy Annotation for CAD Jewellery Images’, Int. J. of Emerging Trends & Technology in Computer Science,Vol. 1, Issue 4. ISSN 2278-6856.

[14] Arunasakthi. K and Kamatchipriya. L(2014), ‘A Review On Linear And Non-Linear Dimensionality Reduction Techniques’, Machine Learning And Applications: An Int. J. (Mlaij), Vol.1, No.1, Pp.65-76.

[15] Julie M. David, Kannan Balakrishnan, (2014), Learning Disability Prediction Tool using ANN and ANFIS, International Journal of Soft Computing, Springer Verlag Berlin Heidelberg, ISSN 1432-7643 (online), ISSN 1433-7479 (print), DOI: 10.1007/s00500-013-1129-0, Vol. 18, Issue 6, pp 1093-1112

[16] Julie M. David, Kannan Balakrishnan, (2012), Attribute Reduction and Missing Value Imputing with ANN: Prediction of Learning Disabilities, International Journal of Neural Computing & Applications, Springer-Verlag London Limited, DOI: 10.1007/s00521-011-0619, Vol. 21, Issue 7, pp 1757-1763

[17] Kresimir Delac, Mislav Grgic and Sonja Grgic (2006), 'Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set', Wiley Periodicals, Inc.

[18] Dr. H.B.Kekre, Sudeep D. Thepade and Akshay Maloo (2010), 'CBIR Feature Vector Dimension Reduction with Eigenvectors of Covariance Matrix using Row, Column and Diagonal Mean Sequences', Int. J. of Computer Applications (0975 – 8887),Vol. 3, No.12.

[19] Sebastian Raschka (2014), 'Linear Discriminant Analysis bit by bit'.

[20] Shereena V.B. and Julie M. David (2014), 'Content Based Image Retrieval : Classification Using Neural Networks', The Int. J. of Multimedia & Its Applications (IJMA) Vol.6, No.5. DOI : 10.5121/ijma.2014.6503 31.

[21] Anca Apatean, Alexandrina Rogozan , Simina Emerich , Abdelaziz Bensrhair, 'Wavelets As Features For Objects Recognition'.

[22] Prof. C. S. Gode, Ms. A. N. Ganar(2014), 'Image Retrieval by Using Color, Texture and Shape Features', Int.Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering, Vol. 3, Issue 4, ISSN (Print) : 2320 – 3765,ISSN (Online): 2278 – 8875.

[23] Hamed Fadaei, and Thotsapon Sortrakul(2014), 'Content-Based Image Retrieval System with Combining Color features and Gradient feature', Int. Conference on Advanced Computational Technologies & Creative Media.

[24] G. Sasikala , R. Kowsalya and Dr. M. Punithavalli (2010), 'A Comparative Study Of Dimension Reduction Techniques For Content-Based Image Retrieval', The Int. J. of Multimedia & Its Applications, Vol.2, No.3, Doi : 10.5121/Ijma.2010.2303 40.

[25] R.Priya, Dr.Vasantha Kalyani David(2012), 'Enhanced Content Based Image Retrieval Using Multiple Feature Fusion Algorithms', Int. J. of Scientific & Engineering Research, Vol. 3, Issue 2, ISSN 2229-5518.

[26] K. Arthi, Mr. J. Vijayaraghavan (2013) 'Content Based Image Retrieval Algorithm Using Color Models', Int. J. of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 3, ISSN (Print) : 2319-5940 ISSN (Online) : 2278-1021.

AUTHORS

Shereena V.B.received her MCA degree from Bharathidasan University, Trichy, India in 2000. During 2000-2004, she was with Mahatma Gandhi University, Kottayam, India as Lecturer in the Department of Computer Applications. Currently she is working as Asst. Professor in the Department of Computer Applications with MES College, Aluva, Cochin, India. She has published papers in International Journals and International and National Conference Proceedings Her research interests include Data Mining, Machine Learning and Image Processing.

Dr. Julie M. David completed her Masters Degree in Computer Applications and Masters of Philosophy in Computer Science in the years 2000 and 2009 in Bharathiyar University, Coimbatore, India and in Vinayaka Missions University, Salem, India respectively. She has also completed her Doctorate in the research area of Artificial Intelligence from Cochin University of Science and Technology, Cochin, India in 2013. During 2000-2007, she was with Mahatma Gandhi University, Kottayam, India, as Lecturer in the Department of Computer Applications. Currently she is working as an Assistant Professor in the Department of Computer Applications with MES College, Aluva, Cochin, India. She has published several papers in International Journals and International and National Conference Proceedings. Her research interests include Artificial Intelligence, Data Mining, and Machine Learning. She is a life member of International Association of Engineers, IAENG Societies of Artificial Intelligence & Data Mining, Computer Society of India, etc. and a Reviewer of Elsevier International Journal of Knowledge Based Systems. Also, she is an Editorial Board Member of various other International Journals. She has coordinated various International and National Conferences.

INTENTIONAL BLANK

TOP K-OPINION DECISIONS RETRIEVAL IN HEALTHCARE SYSTEM

A.Ananda Shankar[1] and Dr.K.R.Ananda Kumar[2]

[1]Associate Professor in Dept. of CSE, Reva University Bangalore, India.
Anishank2003@yahoo.co.in
[2]Professor and HOD Dept. of CSE,S.J.B.I.T, Bangalore, India.
Kra_megha_tn@hotmail.com

ABSTRACT

The aim of this paper is to use data mining technique and opinion mining(OM) concepts to the field of health informatics. The decision making in health informatics involves number of opinions given by the group of medical experts for specific disease in the form of decision based opinions which will be presented in medical database in the form of text. These decision based opinions are then mined from database with the help of mining technique. Text document clustering plays major role in the fast developing information Explosion. It is considered as tool for performing information based operations. Text document clustering generates clusters from whole document collection automatically, normally K-means clustering technique used for text document clustering. In this paper we use Bisecting K-means clustering technique and it is better compared to traditional K-means technique. The objective is to study the revealed groupings of similar opinion-types associated with the likelihood of physicians and medical experts.

KEYWORDS

Opinion Mining (OM),Sentimental Analysis(SA),Data Mining(DM).

1. INTRODUCTION

We live in data-rich times and each day, more data are collected and stored in databases. Increasing the use of data toward answering and understating important questions has driven the development of data mining techniques. The purpose of these techniques is to find information within the large collection of data. Although data mining is a new field of study of medical informatics, the application of analytical techniques to discover patterns has a rich history. Perhaps it was one of the most successful uses of data analysis for discovering and understanding of the medical science, especially infectious disease.

Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician and even on the psycho-physiological condition of the physician. A number of studies have shown that the diagnosis of one patient can differ significantly if the patient is examined by different physicians or even by the same physician at various times . Data mining techniques applied on these databases discover relationships and patterns which are helpful in studying the progression and the management of disease. Prediction

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 57–65, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50906

or early diagnosis of a disease can be kinds of evaluation. About diseases like skin cancer, breast cancer or lung cancer early detection is vital because it can help in saving a patient's life [1].

Healthcare related data mining(DM) is one of the most rewarding and challenging areas of application in data mining and knowledge discovery. The challenges are due to the datasets which are large, complex, heterogeneous, hierarchical, time series and varying of quality. As the available healthcare datasets are fragmented and distributed in nature, thereby making the process of data integration is a highly challenging task.[2]

2. RELATED WORK

Opinion mining (or sentiment analysis(SA)) is the computational study of people's opinions, appraisals, attitudes and emotions toward entities, individuals, issues, events, topics and their attributes. It has become a very active research area in the past few years due to challenging research problems and a wide arrange of applications. There are now at least 40 companies in abroad alone that provide some kinds of opinion mining services. Opinions are important because they are key influences on our behaviors. It is well known that our beliefs and perceptions of reality are to a considerable degree conditioned on how others see the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals, organizations and also for the doctors in medical field. With the explosive growth of social media (i.e., reviews, forum discussions, blogs and social networks, etc) in the past 10 years, individuals and organizations are increasingly using these media for their decision making [3]. For efficiency by reducing system wide costs linked to under treatment, over treatment, by reducing errors, cost and duplication in diagnosis. Nowadays, in medical domain doctors take decision for critical diagnosis with multiple opinion [op]. The literature of U.S healthcare and other sectors tells that 61 % of public and doctors seek for multiple opinions before taking decisions for diagnosis of these diseases [4].

Now days in medical domain, it is difficult to make a decision for complex diseases henceforth doctors seek multiple opinions of different experts in order to achieve the accurate diagnosis process for the diseases [5].

The objective of successful diagnosis is by experts past experience or the knowledge gained from those experience. Experts can make prediction from previous observations (solved cases) and produce diagnosis for new cases. Using these experience experts can suggest good opinions about the diseases. Similarly different experts having different background knowledge (experience) can suggest different opinions, which leads for multiple opinions [6].

3. PROPOSED WORK

The proposed system uses Bisecting K-means algorithm for clustering the common types of opinion decisions given by a set of experts for particular case, where this algorithm will effectively cope up with outliers. In order to retrieve decisions the Best Position algorithm is used.

3.1 Advantages of the Proposed work:

- Bisecting K-means tends to produce clusters of relatively uniform size whereas K-means produce clusters of non-uniform size.
- If the number of clusters is large, then bisecting K-means is more efficient than the Regular K-means algorithm.

- Bisecting K-means is an excellent algorithm for clustering a large number of documents.
- The Bisecting K-means algorithm for document clustering which effectively cope up with outliers.
- Bisecting K-means produces uniform cluster irrespective of centroid selected.
- Best Position algorithm stops early than threshold algorithm and its execution cost Never higher than threshold algorithm.

The Figure (1) below shows flowchart for pre-processing and clustering of documents of opinions. First opinion decisions in the form of text documents are taken as input. In this paper three clusters are used for storing documents. If the number of documents less than number of cluster then the system not going to pre-process and cluster the documents. If the number of documents more than the number of clusters then, the system going to pre-process the documents.

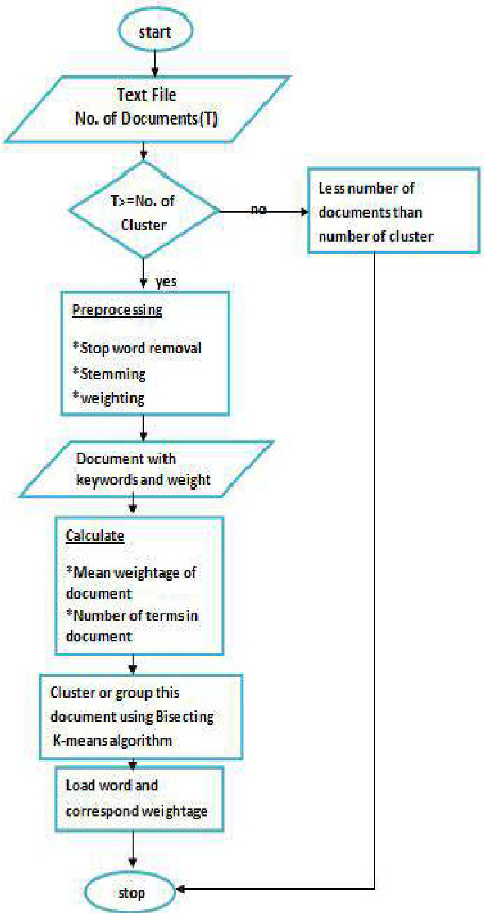

start
Text File
No. of Documents(T)
T>=No. of Cluster
no
Less number of documents than number of cluster
yes
Preprocessing
*Stop word removal
*Stemming
*weighting
Document with keywords and weight
Calculate
*Mean weightage of document
*Number of terms in document
Cluster or group this document using Bisecting K-means algorithm
Load word and correspond weightage
stop

Figure (1).

4. IMPLEMENTATION

The algorithm is implemented in Java and proposed algorithm is based on the clustering of opinion decisions given by the experts for particular case, which is stored in the form of text documents.

4.1. List of Modules

1. Documents pre-processing module.

2. Documents clustering module.

3. Gateway module.

4. User interface module.

4.1.1. Documents pre-processing module:

Document pre-processing is the process of introducing a new document to the information retrieval system in which each document introduced is represented by a set of index terms. The goal of document pre-processing is to represent the documents in such a way that their storage in the system and retrieval from the system are very efficient. Document pre-processing includes the following stages.

Stop word removal:

Stop words are very common words in the natural language. This stop words increase the time of searching particular phrases. Example the, and, who, what, are, was, then etc. Stop words are specific to particular language. It is very difficult to identifying all stop words in a language. We have to identify those stop words manually.

Steps involved in identifying and eliminating stop words:

1. Store the documents contents in the file stream.

2. Manually specify the stop words to remove.

3. Copy the content from file stream to another document.

4. If any stop word encounters don't copy those words to document.

Stemming:

Stemming is to reduce the variant form of words to the normal form. Example connection, connections, connectives, connected, connecting are variant form of the word connect.

Steps involved in identifying and eliminating stem words:

1. Store the documents contents in the file stream.

2. Manually specify the required stemming word.

3. Copy the content from file stream to another document.

4. If any stem word appears before the space character stems those words and copy remaining part to document.

Weighting:

Weighting for the documents done using term frequency inverse document frequency (TF-IDF) weighting function. Number of terms in entire document is document frequency. Number of times the particular word repeated is the term frequency. Weighting is calculated by dividing both measures. Example if document having 1000 words, then the document frequency (DF) is 1000. If a particular word repeated 10 times in document, then the term frequency (TF) is 10. Weight for the particular word is calculated using dividing

(TF) / (DF) that is 10/1000=0.01.

4.1.2. Documents clustering module:

After document pre-processed, all these documents are grouped in three clusters using Bisecting k-means algorithm with the use of two measures mean weight age of all document and mean number of words in all documents.

Bisecting K-means algorithm:

Bisecting K-means tends to produce clusters of relatively uniform size whereas K-means produce clusters of non-uniform size. If the number of clusters is large, then bisecting K-means is more efficient than the regular K-means algorithm. Bisecting K-means is an excellent algorithm for clustering a large number of documents. The Bisecting K-means algorithm for document clustering is effectively coping up with outliers. Bisecting K-means produces uniform cluster irrespective of centroid selected.

Steps followed to clustering documents using Bisecting K-means:

1. Initially put all the documents in a single cluster.

2. Calculate number of words in document and mean weight for documents for all the documents.

3. Calculate the initial centroid by mean number of words in all documents and mean weight age documents.

4. Select the random document and calculate the symmetric point for random documents by considering initial centroid as a midpoint.

5. Find 2 sub-clusters using the basic K-means algorithm.

6. Repeat step 2, the bisecting step, for a fixed number of times and take the split that Produces the clustering with the highest overall similarity. (For each cluster, it similarity is the average pair wise document similarity, and we seek to minimize that sum over all clusters.)

7. Repeat steps 3, 4 and 5 until the desired number of clusters is reached.

4.1.3 Gateway module:

After grouping the documents, Documents action words and weights are transfer to tables. Information extraction is done by gateway module by using the best position algorithm for stopping condition.

Gateway module responsible for the following steps:

1. Receiving the user request.
2. Finding the cluster regarding the request word.
3. Finding the files in cluster belong to the word based on weight age.
4. Sending the request to respective storage node.
5. Receiving the files from storage node.
6. Transfer the file to user system.

Best Position Algorithm (BPA):

Main idea:

Take into account the positions (and scores) of the seen items for stopping Condition and Enables BPA to stop much sooner than Threshold Algorithm (TA).

Best position:

The greatest seen position in a list such that any position before it is also seen. Thus, we are sure that all positions between 1 and best position have been seen.

Stopping condition:

Based on best positions overall score, i.e. the overall score computed based on the best positions in all lists

How the algorithm works:

- Do sorted access in parallel to each list Li, For each data item seen in Li, Do random access to the other lists to retrieve the item's score and position.
- Maintain the positions and scores of the seen data item, Compute best position in Li and Compute best positions overall score.
- Stop when there are at least k data items whose overall score ≥ best positions overall score

4.1.4 User interface module:

First user has to register and login with those registration details. The user has to enter keyword for retrieving documents. The requested keyword passes into Gateway module to retrieve the documents. If documents having the keyword the user searched then documents are retrieved based on highest weight age.

User responsible for the following action:

- User has to register and login with registered details.
- Input search word.
- View the Received file name.

- Open the file content and view.

5. TEST CASES :

5.1. Test Cases for Unit Testing:

Serial # Test Case	UTC-1
Name of Test	Pre-process button in User Interface Window
Items being tested	Pre-Process Button
Sample Input	Click on Pre-process Button
Expected Output	Files in the input folder should be weighted and stemmed.
Actual output	Files in the input folder are processed according to the expectation.
Remarks	Pass

5.2. Efficient Top-k opinion decision retrieval Testing and validation:

Serial # Test Case	UTC-2
Name of Test	Bisecting K-means Button in User Interface window.
Items being tested	Bisecting K-Means Button
Sample Input	Click on Bisecting K-Means Button.
Expected Output	After clicking the button clustering operation should performed on those files.
Actual output	Clustering operation is happened after clicking the button.
Remarks	Pass

5.3. Efficient Top-k opinion decision retrieval Testing and validation:

Serial # Test Case	UTC-3
Name of Test	User Home-Submit
Items being tested	One word in text field with submit button
Sample Input	Click on submit Button
Expected Output	It will check the files which are found with the worked what you are typing. If found then it should display the file downloaded message. In Gateway frame should display the selected files with highest weight.
Actual output	It is displaying.
Remarks	Pass

5.4. Test Cases for Integration Testing:

Serial # Test Case	IT
Name of Test	Clustering Documents
Description	It should cluster the documents and stores in respective cluster
Sample Input	Pre-Processed Documents.
Expected Output	Documents should be stored in respective cluster.
Actual output	Documents are stored in respective cluster based on Bisecting K-Means.
Remarks	Pass

5.5. Checking the overall working of system:

Serial # Test Case	ST
Name of Test	Checking the overall working of system
Description	It will retrieve the highest weighted documents.
Sample Input	Text documents.
Expected Output	Documents having requested keyword with highest weight.
Actual output	It is retrieving documents having requested keyword with highest weight.
Remarks	Pass

6. CONCLUSION

In this work, propose a novel approach that carefully uses text mining techniques in order to pre-process and clustering text documents. Documents pre-processing is done order to reduce the time of information extraction. Normally K-means clustering technique used for text document clustering. In this work uses Bisecting K-means clustering technique and it is better compared to traditional K-means technique. Further this clustered document used for performing information based operations. This paper concentrates on finding Top-K file retrieval based on keyword weight in document. The best position algorithm (BPA) which executes top-k queries more efficiently than Threshold Algorithm (TA). BPA stops as early as TA, and that its execution cost is never higher than TA.

7. FUTURE ENHANCEMENT

The future work is pursuing in the following direction: The paper can be extended to deal with still more formats of documents.

REFERENCES

[1] Ashwin Kumar.U.M, Ananda Kumar.K.R. " A web based patient support system using Artificial Intelligence to improve Health Monitoring and Quality of life " in Advanced Computing and Communication Technologiea(ACCT)-2012 Second International conference on 7-8-Jan2012,ISBN 978-1-4673-0471-9.

[2] Atul Kumar Pandey*, Prabhat Pandey**, K.L. Jaiswal***, Ashish Kumar Sen**** " DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute election Method" ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, October 2013.

[3] Yuefeng li, Ning Zhong, Raymond Y.K.lau " Topic Feature Discovery and Opinion Mining"10thIEEEinternational Conference on Data Mining(ICDM'10. Sydney, Australia.

[4] James Manyika,Michael Chui,Brad Brown,Jacques Bughin, Richard Dobbs, Charles Roxburgh Angela Hung Byers from McKinsey Global Institute " Big data: The next frontier for innovation, competition, and productivity " McKinsey & Company 2011.

[5] Mitja Lenic, Petra Povalej, Milan Zorman, Peter Kokol, faculty of electrical engineering and computer science, university of Maribor, slovenia,Proceedings of the 17th IEEE symposium on computer-based medical systems(CBMS'04) 1063- 7125/04,2004 IEEE.

[6] Mitja Lenic, Petra Povalej, Milan Zorman, Peter Kokol, faculty of electrical engineering and computer science, university of Maribor, slovenia, Proceedings of the 17th IEEE symposium on computer-based medical systems(CBMS'04) 1063- 7125/04,2004 IEEE.

AUTHORS

A. Ananda Shankar[1] is a M.Tech graduate in Computer Science and Engineering. Currently works as Associate Professor in the School of Computing and Information Technology at Reva University, Bangalore, Karnataka. Currently he is doing his research work in the area of Medical Data mining under Visvesvaraya Technological University, Belgaum. He has Two international conference publications and One international journal publications and Two national conference publication to his credit.

K. R. Ananda Kumar[2] holds a Doctoral Degree in Computer Science and Engineering. Currently works as Professor and Head of the Department in the Department of Computer Science and Engineering, S.J.B.I.T, Bangalore, Karnataka. He has a vast teaching experience of about 25 years. His research interest includes medical data mining; data stream mining, Artificial intelligence, intelligent agents and web mining. He is currently guiding five research scholars. He has Over 51 papers in International & National Journals and conferences to his credit.

INTENTIONAL BLANK

ANALYSIS OF SUPPLIER'S PERFORMANCE THROUGH FPIR/FNIR AND MEMBERSHIP DEGREE TRANSFORMATION

S.Hemalatha[1] K. Ram Babu[2] K.Narayana Rao[3] K.Venkatasubbaiah[4]

[1]Department of Mechanical Engg, Lendi Institute of Engineering and Technology, Vizianagaram- 535 005
[2]Department of Mechanical Engg, Andhra University, Visakhapatnam-530 007
[3]Government Model Residential Polytechnic, Paderu-531024
[4]Department of Mechanical Engg, Andhra University, Visakhapatnam-530 007

ABSTRACT

In today's highly competitive business environment, evaluation of suppliers is the prime function of the purchasing department of the organization. It is due to the fact that high percentage of the material cost for manufacturing of a product is involved. Identification of decision criteria and methods for supplier evaluation are appearing to be the important research area in the literature. In this paper, hybrid methodology of Fuzzy positive Ideal rating /Fuzzy Negative Ideal rating and Membership Degree Transformation- M (1, 2, 3) is proposed for evaluation of supplier's performance. A wide literature review is made and six selection criteria namely: Cost, Quality, Service, Business performance, Technical Capability and Delivery performance are considered for evaluation. A detailed application of the proposed methodology is illustrated. The proposed methodology is useful not only to judge the overall performance of the supplier but also to know which criteria/sub-criteria need to be improved.

KEYWORDS

Membership Degree transformation; Fuzzy positive Ideal Rating; Fuzzy Negative Ideal Rating; Supplier performance;

1. INTRODUCTION

The traditional business functions need to be coordinated to achieve customer satisfaction, value, profitability, and competitive advantage for individual companies and the entire supply chain. One of the functions that have been singled out as important in the coordination processes of the individual firms and supply chain is purchasing.

Cheraghi et al (2002) presented the critical success factors (CSFs) for supplier selection reported in the literature emanating from the seminal work of Dickson (1966) and provide an update based on reviewing more than 110 research papers. The authors indicated significant change in the relative importance of various critical success factors in the research reported during 1966-1990 versus 1990-2001. Supplier selection and their performance evaluation is one of the important

Dhinaharan Nagamalai et al. (Eds) : CSTY, SIGI, AI, MaVaS - 2015
pp. 67–79, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50907

drivers of supply chain performance. Uses of suitable criteria with appropriate methodologies are necessary for performance evaluation of a supplier. In the literature, it is observed that supplier selection and evaluation methods were based on quoted price, quality, business relations, lead time etc., constitute a multi-criteria or multi-objective decision making problem. The overall objective of the supplier selection process is to identify, evaluate, contract with the suppliers and optimum quota allocation to the suppliers. Boer et al (2001) made a review on decision methods on supplier selection based on academic literature. Byun (2001) presented Analytical Hierarch Process (AHP) approach for vendor selection and identified supplier reliability, product quality and supplier experiences are the critical factors for effective supplier selection in Korean automobiles. Muralidharan et al (2002) suggested guidelines for comparing supplier attributes using a five-point rating scale and developed aggregation technique for combining group member's preferences into one consensus for supplier rating. In the supplier selection process, organizations judge the supplier's ability to meet the requirements of the organization to survive in the intensely competitive global economy. Dulmin and Mininno (2003) used multi-criteria decision analysis method in supplier selection problem using PROMETHEE and GAIA methodology. Rajkumar and Ray (2004) identified attributes and factors relevant for performance evaluation of suppliers through fuzzy inference system of the MATLAB fuzzy logic tool box. Venkatasubbaiah and Narayana Rao (2004) considered thirty three sub-criteria under six main criteria reported in the literature in four decision hierarchy levels for supplier selection using AHP. Very often, experts opinion is the prominent characteristic of multi-criteria decision making problems and this impreciseness of human's judgments can be handled through the fuzzy sets theory developed by Zadeh (1965). Fuzzy set theory effectively incorporates imprecision and subjectivity into the model formulation and solution process. Chen et al (2006) adopted TOPSIS concept in fuzzy environment to incorporate imprecision and subjectivity into the model formulation and solution process to determine the ranking order of the suppliers. The author considered the factors such as quality, price, and flexibility and delivery performance. Lee et al (2007) adopted Fuzzy Analytic Hierarchy Process (FAHP) to analyze the importance of multiple factors by incorporating the experts' opinions to select Thin Film Transistor Liquid Crystal Display (TFT-LCD) suppliers. Narayana Rao et al (2007) illustrated fuzzy outranking technique for selection of supplier using minimum and gamma operators for aggregating the concordance and discordance indices of the alternative suppliers to arrive the ranking of suppliers with credibility values. Shouhua Yuan et al (2008) proposed DEA, AHP and fuzzy set theory to evaluate the overall performance of suppliers of a manufacturing company. Enyinda et al (2010) adopted analytic hierarchy process (AHP) model and implemented using Expert Choice Software for a supplier selection problem in a generic pharmaceutical organization. Elanchezhian et al (2010) adopted analytical network process (ANP) and TOPSIS method for select the best vendor. Jitendra Kumar and Nirjhar Roy (2010), adopted a hybrid model using analytic hierarchy process (AHP) and neural networks (NNs) theory to assess vendor performance. Yucel and Guneri (2011) assessed the supplier selection factors through fuzzy positive ideal rating and negative ideal rating to handle ambiguity and fuzziness in supplier selection problem and developed a new weighted additive fuzzy programming approach. Yang and Jiang (2012) proposed AHM (Analytic Hierarchy Method) and M(1,2,3) methodology to evaluate the supply chains' overall performance. Durga Prasad et al (2012) proposed and illustrated the methodology for evaluating the efficiency and performance of the suppliers using Data Envelopment Analysis (DEA) technique. Amindoust (2012) proposed and illustrated ranking methodology in fuzzy environment with sustainable supplier selection criteria/sub-criteria. Abbasi et al (2013) proposed a framework and applied QFD/ANP to rank the relative importance of the key attributes in selection of suppliers. Galankashi et al (2013) presented supplier Selection for Electrical Manufacturing Companies Based on Different Supply Chain Strategies using AHP. Eshtehardian et al (2013), presented a decision support system to the supplier selection in the construction and civil engineering companies using AHP and ANP simultaneously. Om pal et al (2013) presented review on supplier selection criteria and methods basing on research reported in the supply chain management area. Deshmukh and Vasudevan (2014) explored criteria that are important for green supplier selection, as evident in literature and gathered from discussions with experts. Ergün and Atalay (2014) proposed FAHP and FTOPSIS for evaluation of suppliers of an electronic company.

From the review of literature, it is observed that there is limited research in group decision approach for prioritizing the supplier selection criteria in fuzzy environment. Further, classification of a supplier belongs to a particular class basing on the data mining technology is also limited. In lieu of this, a hybrid methodology is proposed for evaluation of supplier's performance and illustrated by considering the supplier of a pharmaceutical company. In the methodology, Fuzzy positive Ideal Rating and Fuzzy Negative Ideal rating approach is adopted to find out the importance weights of criteria/sub-criteria. Then, Membership transformation method – M(1,2,3) is adopted to find out the grade of overall performance of a supplier. Proposed methodology is explained in section two. Numerical Illustration is presented in section three. Results and discussion is made in section four. Finally, the conclusions are summarized with future scope in section five.

2. METHODOLOGY

Step 1: Establish Evaluation Index System of Supplier Performance

An Organization has to identify criteria for supplier selection to evaluate whether the supplier fits its competitive strategy and supply chain strategy .The total performance of the supplier depends on the capabilities in each criteria/sub criteria and the relative importance given to them.

Step 2: Determine importance weights of the criteria/sub criteria

Fuzzy Positive Ideal Rating (FPIR) and Fuzzy Negative Ideal Rating (FNIR) are used to compute the weights of the criteria/sub criteria (Yucel and Guneri, 2011).

Step 3: Membership Transformation through "Effective, Comparison and Composition"

Membership transformation method – M(1,2,3) proposed by Hua and Ruan (2009) as discussed in the following steps is adopted to determine the evaluation matrix of the alternative.

Step 3.1: Determine Evaluation Membership $\mu_{jk}(Q)$

Percentage of satisfaction among the domain experts under each class is considered as evaluation matrix of each criterion.

$\mu_{jk}(Q)$=membership of jth sub-criteria of the criteria group 'Q' belonging to the kth fuzzy membership class.

Step 3.2: Determine Distinguishable Weights ($\alpha_j(Q)$)

Distinguishable weight represents the normalized and quantized value obtained from the following relation.

$$\alpha_j(Q) = v_j(Q) / \sum_{j=1}^{m} v_j(Q) \qquad (j = 1..m)$$

Where

$$v_j(Q) = 1 - (1/\log(p)) * H_j(Q)$$

$$H_j(Q) = -\sum_{k=1}^{p} \mu_{jk}(Q) * \log \mu_{jk}(Q)$$

$v_j(Q)$ = weight of the jth sub criteria of the evaluation criteria object ‘Q’ obtained from uncertainty in the payoff information of the sub criteria

$H_j(Q)$ = Measure of uncertainty in the payoff information of the jth sub criteria of the evaluation criteria object ‘Q’

Step 3.3: Determine Comparable sum Vector $M_k(Q)$

Comparable value of the sub criteria under the given criteria is determined from the following relation

$$M_k(Q) = \sum_{j=1}^{m} \beta_j(Q) * \alpha_j(Q) * \mu_{jk}(Q)$$

$\beta_j(Q)$ = Importance Weight Vector of sub-criteria

Step 3.4: Determine Membership Vector $\mu_k(Q)$

Membership vector of the object ‘Q’ belonging to class ‘k’ is determined from the following relation.

$$\mu_k(Q) = M_k(Q) / \sum_{k=1}^{p} M_k(Q)$$

Step 3.5: Determine Evaluation Matrix of the alternative U(S)

Membership matrix of all the criteria of the object ‘Q’ is determined and evaluation matrix is formed as shown below.

$$U(S)=\begin{pmatrix}\mu(C1)\\ \mu(C2)\\ \mu(C3)\\ \mu(C4)\\ ..\\ ..\end{pmatrix}$$

Step 4: Determine Final membership Vector $\mu(S)$

Once the evaluation matrix of the goal and the weights of the each criterion are known the procedure is repeated from the step 3.1 to 3.5 is repeated to obtain the final membership vector of the goal.

Step 5: Determine the grade of overall Performance (K_O)

Overall performance of the alternative is determined by applying confidence recognition rule (Confidence degree: λ >0.7)

$$K_O = \min\ \{k| \sum_{k=1}^{k} \mu_k(S) \geq \lambda\ \}$$

3. NUMERICAL ILLUSTRATION

In this paper, supplier performance evaluation using proposed methodology is illustrated with a numerical example. Supplier's performance metrics taken from the literature (Venkatasubbaiah et al., 2004; Lee et al., 2007; Narayana rao et al., 2007) are considered for performance evaluation of supplier. The evaluation hierarchy is organized into three layers namely, Goal, Criterion layer and sub-criterion layer as shown in fig 1.

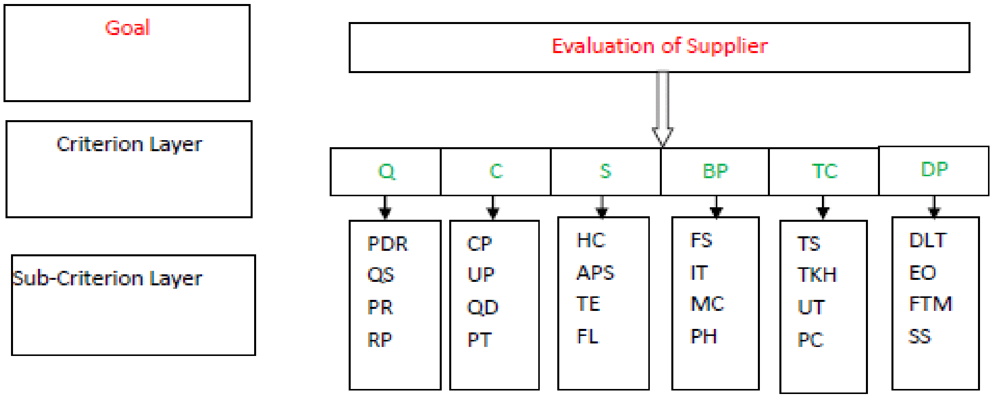


Figure 1: Hierarchy of Evaluation Index System of Supplier performance

Evaluation of supplier performance is considered as goal. Supplier evaluation criteria namely, Quality (Q), Cost (C), Service (S), Business performance (BP), Technical Capability (TC) and Delivery are considered at criterion level. Sub-criteria under each criterion are given below.
Sub-criteria under Quality (Q): Product durability and Reliability (PDR); Quality systems (QS); Percent Rejection (PR); Reputation and Position in the market (RP);

Sub-criteria under Cost (C): Competitive Pricing (CP); Unit Price (UP); Quantity Discount (QD); Payment Terms (PT);

Sub-criteria under Service (S): Handling of Complaints (HC); Availability of product/service (APS); Training equipment (TE); Flexibility (FL);

Sub-criteria under Business Performance (BP): Financial Stability (FS); IT usage (IT); Management Capability (MC); Personnel Capability (PC);

Sub-criteria under Technical Capability (TC): Technical Support (TS); Technology Know How (TKH); Understanding of Technology (UT); Performance History (PH);

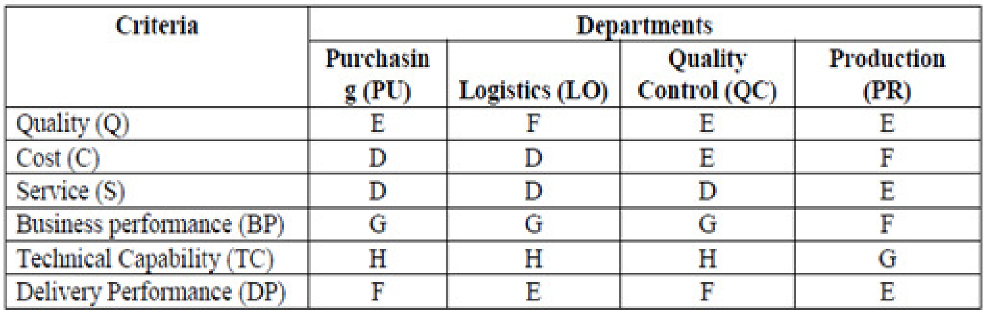
3.6 Sub-criteria under Delivery Performance (DP): Delivery of Lead Time (DLT); Expeditation of Orders (EO); Flexible Transportation Modes (FTM); Safety and Security of Components (SS);
Necessary data on the relative importance of criteria/sub-criteria gathered from discussions with the managers of Purchasing, Logistics, Quality Control and Production departments of a pharmaceutical company. These industries need to improve their supply chain performance by concentrating on supplier issues to face with the uncertainty within the business environment.

3.1 Relative weights of the criteria/sub-criteria

Relative weights of criteria/sub-criteria are determined as discussed in step 2 of the methodology section. Data is collected by discussion with the managers of Purchasing, Logistics, Quality Control and Production departments to assess the relative importance of the criteria on the supplier performance. Degree of relative importance of criteria is presented with the linguistic variables: Nill-NL;Very Low- VL; Low-L; Medium Low- ML; M- Medium; Medium High- MH; High- H;Very High- VH; Full- F; Aggregated responses of the importance of criteria and sub-criteria in terms of the linguistic variables by the employees of different departments are shown in the table 1.

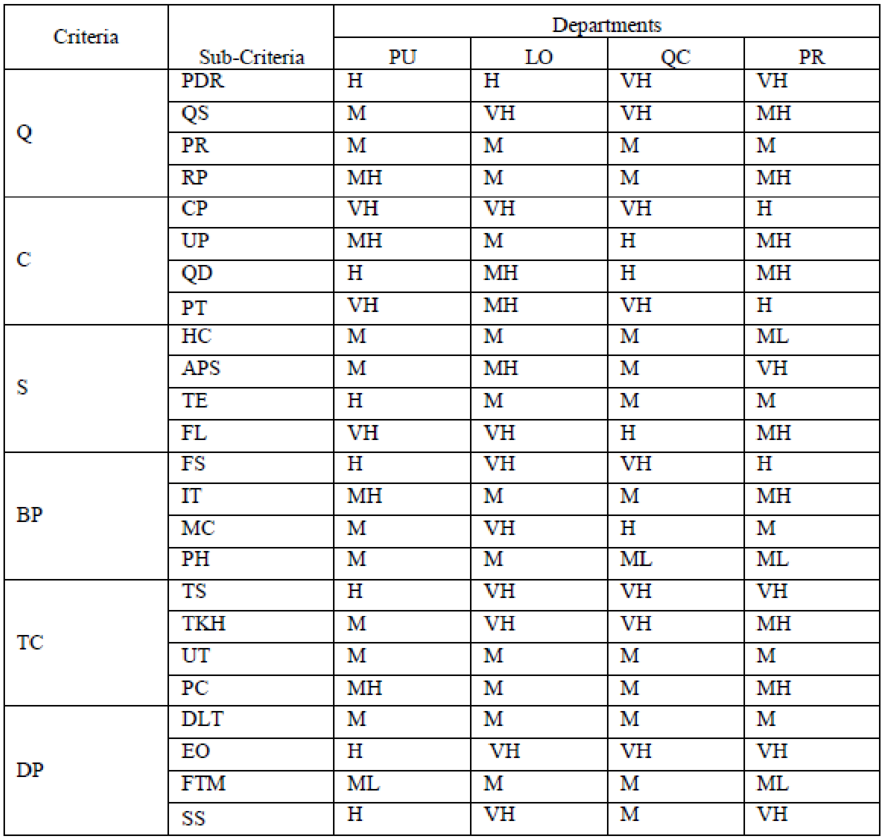
The study considered the above criteria/sub-criteria from the literature and these are prioritized. Relative weights of criteria and sub-criteria are determined from the aggregated responses shown table 1 and table 2 respectively through Fuzzy Positive Ideal Rating (FPIR) and Fuzzy Negative Ideal Rating (FNIR) approach as discussed in step 2 of the methodology section. Relative weights of criteria and sub-criteria are shown in table 3.From table 3 it is observed that Quality criterion is highly prioritized followed by Technical capability, Delivery Performance, Cost, Service and Business Performance. This is due to the fact that the pharmaceutical company considers Quality is the most important criterion that must be evaluated for successful selection of the supplier. Technical Capability criterion is ranked second since it is an obvious consideration for any pharmaceutical company. Relative weights of the criteria/sub-criteria are shown in fig 2

Table 1: Aggregated Responses on Criteria



Criteria	Departments			
	Purchasing (PU)	Logistics (LO)	Quality Control (QC)	Production (PR)
Quality (Q)	E	F	E	E
Cost (C)	D	D	E	F
Service (S)	D	D	D	E
Business performance (BP)	G	G	G	F
Technical Capability (TC)	H	H	H	G
Delivery Performance (DP)	F	E	F	E

Table 2: Aggregated Responses of Sub-criteria



Criteria	Sub-Criteria	Departments			
		PU	LO	QC	PR
Q	PDR	H	H	VH	VH
	QS	M	VH	VH	MH
	PR	M	M	M	M
	RP	MH	M	M	MH
C	CP	VH	VH	VH	H
	UP	MH	M	H	MH
	QD	H	MH	H	MH
	PT	VH	MH	VH	H
S	HC	M	M	M	ML
	APS	M	MH	M	VH
	TE	H	M	M	M
	FL	VH	VH	H	MH
BP	FS	H	VH	VH	H
	IT	MH	M	M	MH
	MC	M	VH	H	M
	PH	M	M	ML	ML
TC	TS	H	VH	VH	VH
	TKH	M	VH	VH	MH
	UT	M	M	M	M
	PC	MH	M	M	MH
DP	DLT	M	M	M	M
	EO	H	VH	VH	VH
	FTM	ML	M	M	ML
	SS	H	VH	M	VH

Table 3: Relative weights of the criteria/sub criteria

Criteria	Weight	Sub-criteria	Weight	Criteria	Weight	Sub-criteria	Weight
C	0.1542	CP	0.2787	TC	0.2025	TS	0.3093
		UP	0.214			TKH	0.2555
		QD	0.2431			UT	0.2029
		PT	0.2642			PC	0.2323
S	0.1388	HC	0.1941	Q	0.2219	PDR	0.2585
		APS	0.3019			QS	0.2238
		TE	0.2361			PR	0.2053
		FL	0.2679			PC	0.3124
BP	0.1239	FS	0.3148	DP	0.1587	DLT	0.2105
		IT	0.2282			EO	0.3244
		MC	0.2553			FTM	0.1878
		PH	0.2017			SS	0.2774

Table 4: Evaluation Responses and Memberships

Criteria	Sub-Criteria	Evaluation Responses					Evaluation Memberships				
		VS	SA	GE	DS	VD	VS	SA	GE	DS	VD
C	CP	13	12	20	19	11	0.1733	0.1600	0.2667	0.2533	0.1467
	UP	14	9	32	14	6	0.1867	0.1200	0.4267	0.1867	0.0800
	QD	11	17	23	14	10	0.1467	0.2267	0.3067	0.1867	0.1333
	PT	9	10	30	17	9	0.1200	0.1333	0.4000	0.2267	0.1200
S	HC	13	17	20	18	7	0.1733	0.2267	0.2667	0.2400	0.0933
	APS	13	12	30	9	11	0.1733	0.1600	0.4000	0.1200	0.1467
	TE	12	8	32	11	12	0.1600	0.1067	0.4267	0.1467	0.1600
	FL	11	15	26	12	11	0.1467	0.2000	0.3467	0.1600	0.1467
BP	FS	13	12	25	16	9	0.1733	0.1600	0.3333	0.2133	0.1200
	IT	9	24	23	11	8	0.1200	0.3200	0.3067	0.1467	0.1067
	MC	27	14	15	12	7	0.3600	0.1867	0.2000	0.1600	0.0933
	PH	24	17	16	6	12	0.3200	0.2267	0.2133	0.0800	0.1600
TC	TS	23	17	13	12	10	0.3067	0.2267	0.1733	0.1600	0.1333
	TKH	20	20	16	12	7	0.2667	0.2667	0.2133	0.1600	0.0933
	UT	30	10	15	7	13	0.4000	0.1333	0.2000	0.0933	0.1733
	PC	21	20	10	18	6	0.2800	0.2667	0.1333	0.2400	0.0800
Q	PDR	25	17	6	15	12	0.3333	0.2267	0.0800	0.2000	0.1600
	QS	18	16	15	13	13	0.2400	0.2133	0.2000	0.1733	0.1733
	PR	15	10	25	11	14	0.2000	0.1333	0.3333	0.1467	0.1867
	PC	15	13	28	12	7	0.2000	0.1733	0.3733	0.1600	0.0933
DP	DLT	21	12	25	10	7	0.2800	0.1600	0.3333	0.1333	0.0933
	EO	19	7	27	13	9	0.2533	0.0933	0.3600	0.1733	0.1200
	FTM	13	15	22	13	12	0.1733	0.2000	0.2933	0.1733	0.1600
	SS	16	9	26	11	13	0.2133	0.1200	0.3467	0.1467	0.1733

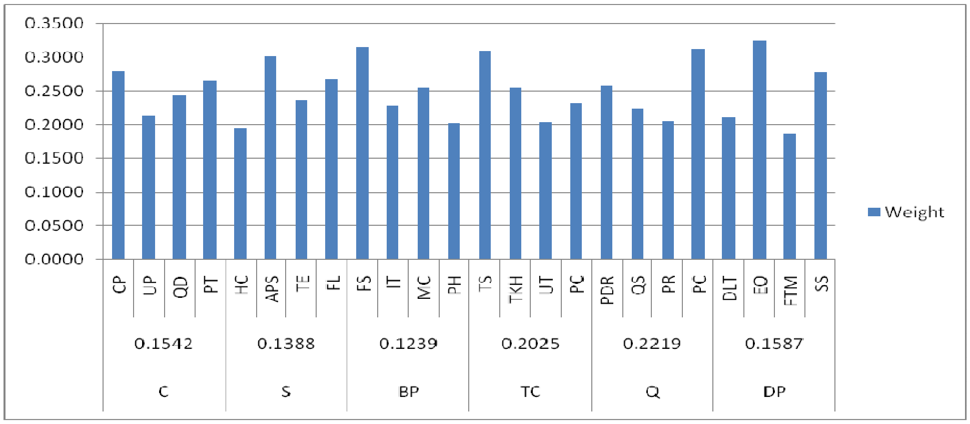


Figure 2: Relative weights of the criteria/sub-criteria

3.2 Evaluation Membership

Data on the given supplier performance sub-criteria is obtained from 75 employees of production, Logistics, Quality control and Marketing & sales departments of the pharmaceutical company. No of employees responded regarding the satisfaction levels in five classes and the membership values are shown in table 4.

3.3 Evaluation matrix

Evaluation Matrix is determined as discussed in step 3 of methodology section. Evaluation matrix of supplier's performance is shown below.

$$\text{U(S)} = \begin{pmatrix} \mu(C) \\ \mu(S) \\ \mu(BP) \\ \mu(TC) \\ \mu(Q) \\ \mu(DP) \end{pmatrix} = \begin{pmatrix} 0.1530 & 0.1400 & 0.3896 & 0.2093 & 0.1081 \\ 0.1632 & 0.1583 & 0.3779 & 0.1576 & 0.1429 \\ 0.2365 & 0.2420 & 0.2606 & 0.1422 & 0.1187 \\ 0.3324 & 0.2039 & 0.1825 & 0.1503 & 0.1309 \\ 0.2581 & 0.1901 & 0.2357 & 0.1751 & 0.1411 \\ 0.2458 & 0.1234 & 0.3459 & 0.1566 & 0.1283 \end{pmatrix}$$

3.4 Final membership Vector

Final membership vector of the supplier's performance is determined as discussed in step 4 of the methodology section. The Final membership vector of the supplier's performance is shown below.

$$\mu(S) = \begin{pmatrix} 0.2218 & 0.1639 & 0.3182 & 0.1699 & 0.1262 \end{pmatrix}$$

3.5 Grade of Overall Performance of the supplier

From the numerical illustration, according to the final membership vector, it is observed that the overall performance of the supplier belongs to the 'General' level with the confidence level of 70.39% (22.18%+16.39%+31.82%).

4. RESULTS AND DISCUSSION

Evaluation membership of supplier's performance is shown in fig 2. From the figure, it is understood that Technical Capability (TC) of the supplier is showing relatively high confidence level of performances of 33.24% in 'Very Satisfied' level. Cost (C), Service (S), Supplier performance in respect of Business performance (BP), Technical Capability (TC), Quality (Q), and Delivery Performance (DP) are showing confidence levels of 38.96%, 37.79%, 26.06%, 18.25%, 23.57% and 34.59% respectively in 'General' level.

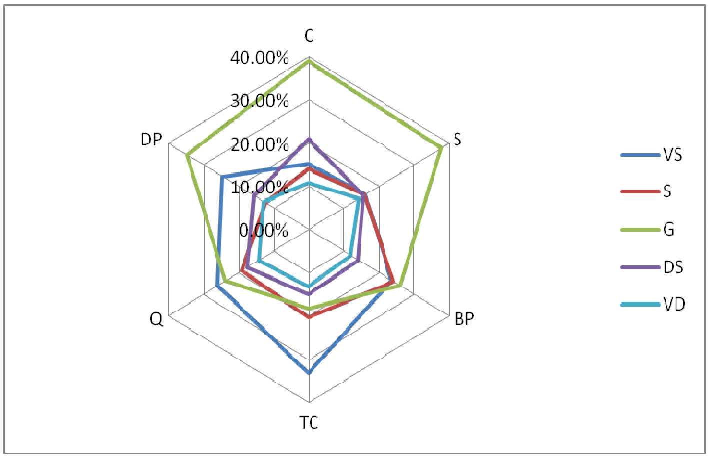


Figure 2: Evaluation memberships of supplier's performance criteria

From the results of the final membership values, it can be judged that the performance of the supplier is considered as 'General' level as the obtained confidence level (70.39%) is more than the minimum confidence level of 70%. Overall confidence level with 'Very Satisfied' is only 22.18% indicates that the supplier should improve the performance from every criteria. In the context of supplier evaluation for a pharmaceutical company, the suppliers need to improve quality, technical capability and delivery performance such that the purchasing company will be capable of rapidly responding to changes to their customer demands. Implementing continuous quality improvement methods, making use of latest equipments and machines, implementing new thoughts in business processes will be useful to improve the supplier's performance

5. CONCLUSIONS

The proposed methodology is a hybrid methodology that combined the FPIR/FNIR approach with Membership transformation method – M (1,2,3) to evaluate the performance of supplier. The proposed methodology is useful not only to judge the overall performance of the supplier but also to know which criteria/sub-criteria need to be increased. The proposed hybrid method is useful to evaluate the supplier's performance as it is affected by the subjective judgment involved in measuring of the criteria/sub-criteria by the stake holders. The methodology maybe extended for the supplier evaluation and selection basing on the supply chain strategy (Lean, Agile and Leagile). To this effect, it requires critical judgment to assess the relative weights among the criteria basing on lean, agile and leagile supply chain strategies. Also, the study can be extended to other areas of decision making in evaluation and ranking of alternatives. Also, the performance of the proposed method can be improved by reducing the subjective judgment in prioritizing the factors/sub-factors.

ACKNOWLEDGEMENTS

The authors are very much thankful to the reviewer for making constructive comments to improve the quality of the paper.

REFERENCES

[1] Abbasi.M., R. Hosnavi, and B. Tabrizi, (2013) "An Integrated Structure for Supplier Selection and Configuration of Knowledge-Based Networks Using QFD, ANP and Mixed-Integer Programming Model", Journal of Industrial Engineering, pp.1-8

[2] Amindoust Atefeh, Shamsuddin Ahmeda, Ali Saghafinia and Ardeshir Bahreininejada, (2012) " Sustainable supplier selection: A ranking model based on fuzzy inference system", Applied Soft Computing, Vol.12, No.6, pp.1668–1677

[3] Boer. Luitzen de, Eva Labro and Pierangela Morlacchi, (2001) " A review of methods supporting supplier selection", European Journal of Purchasing & Supply Management, Vol. 7, pp.75-89

[4] Byun. D., (2001) "The AHP Approach for Selecting an Automobile Purchase model", Information and Management, Vol. 38, pp. 289–297.

[5] Chen-Tung Chena, Ching-Torng Linb and Sue-Fn Huang, (2006) "Fuzzy approach for supplier evaluation and selection in supply chain management", International Journal of Production Economics, Vol.102, No.2, pp.289-301

[6] Cheraghi. S. Hossein, Mohammad Dadashzadeh and Muthu Subramanian, (2002) " Critical Success factors For Supplier Selection: An Update", Journal of Applied Business Research, Vol. 20, N0.2, pp.93-108

[7] Dickson, G.W., (1966) "An analysis of vendor selection systems and decisions", Journal of Purchasing, Volume 2, No.1, pp. 5

[8] Deshmukh. Ashish. J and Hari Vasudevan, (2014) "Emerging Supplier Selection Criteria in the Context of Traditional Vs Green Supply Chain Management", International Journal of Managing Value and Supply Chains,Vol.5, No. 1, pp.19-33

[9] Dulmin Riccardo and Valeria Mininno (2003), "Supplier selection using a multi-criteria decision aid method", Journal of Purchasing and Supply Management Vol. 9, No. 4, pp.177-187

[10] Durga Prasad.K.G , K.Venkata Subbaiah, Ch. Venu Gopala Rao and K.Narayana Rao, (2012) " Supplier Evaluation Through Data Envelopment Analysis", Journal of Supply Chain Management Systems, Vol.1, No.2, pp.1-11

[11] Elanchezhian.C., B. Vijaya Ramnath and R. Kesavan, (2010) " Vendor Evaluation Using Multi Criteria Decision Making Technique", International Journal of Computer Applications, Vol. 5, N0.9, pp.4-9

[12] Enyinda, Chris I., Emeka Dunu and Fesseha Gebremikael, (2010) "An Analysis of Strategic Supplier Selection and Evaluation in a Generic Pharmaceutical Firm Supply Chain", Proceedings of ASBBS, Los Vegas, February 2010, Vol.17, No.1, pp.77-91

[13] Ergün Eraslan and Kumru Didem Atalay, (2014) "A Comparative Holistic Fuzzy Approach for Evaluation of the Chain Performance of Suppliers", Journal of Applied Mathematics, pp.1-9

[14] Eshtehardian Ehsan, Parviz Ghodousi and Azadeh Bejanpour, (2013) "Using ANP and AHP for the Supplier Selection in the Construction and Civil Engineering Companies; Case Study of Iranian Company", KSCE Journal of Civil Engineering, Vol.17, No.2, pp.262-270

[15] Galankashi Masoud Rahiminezhad, Anoosh Moazzami, Najmeh Madadi, Arousha Haghighian Roudsari and Syed Ahmad Helmi, (2013) "Supplier Selection for Electrical Manufacturing Companies Based on Different Supply Chain Strategies", International Journal of Technology Innovations and Research, pp.1-13

[16] Hua Jiang and Junhu Ruan, (2009) "Fuzzy Evaluation on Network security based on the New Algorithm of Membership Degree Transformation- M(1,2,3)" , Journal of Networks, Vol. 4, No.5, pp. 324-331

[17] Jitendra Kumar and Nirjhar Roy, (2010) "A Hybrid Method for Vendor Selection using Neural Network", International Journal of Computer Applications, Vol. 11, No. 2, pp.35-40

[18] Lee. A.H., H.Y.Kans, E-M.Lai, W.M.Way and C.F.Hou, (2007) "TFT-LCD supplier selection by Poun stream manufacture using fuzzy Multi-choice Goal Programming", Proceeding of Computational Intelligence Conference, Banff, Aberta, Canada, pp. 574.

[19] Muralidharan.C, N. Anantharaman., S.G. Deshmukh., (2002) "A multi-criteria group decision making model for supplier rating", The Journal of Supply Chain Management, Vol.38, No.4, pp.22 – 23

[20] Narayana Rao. K., K.Venkata subbaiah, V. Rama Chandra Raju, (2007) "Supplier Selection in Supply Chain Management through Fuzzy Outranking Technique", Industrial Engineering, Vol.XXXVI, No.09, pp.17-21.

[21] Om Pal, Amit Kumar Gupta and R. K. Garg, (2013) "Supplier Selection Criteria and Methods in Supply Chains: A Review", International Journal of Social, Education, Economics and Management Engineering, Vol.7, No.10, pp.1395-1401

[22] Rajkumar Ohdar and Pradip kumar Ray, (2004) "Performance measurement and evaluation of suppliers in supply chain in evalutionary fuzzy based approach", Journal of Manufacturing Technology Management, Vol.15, No.8, pp. 723 – 734

[23] Shouhua Yuan, Xiao Liu, Yiliu Tu and Deyi Xue, (2008) "Evaluating Supplier Performance Using DEA and Piecewise Triangular Fuzzy AHP", Journal of Computing and Information Science in Engineering,Vol.8, pp.1-7

[24] Venkata Subbaiah. K., Narayana Rao. K., (2004) "Supplier selection in Supply Chain Management through AHP", Proceedings of VIII Annual International Conference, The Society of Operations Management, Mumbai, pp.72-80

[25] Yang Jing and Hua Jiang, (2012) “Fuzzy Evaluation on Supply Chains’ Overall Performance Based on AHM and M (1,2,3)”, Journal of Software , Vol.7, No.12, pp. 2779-2786

[26] Yucel Atakan and Ali Fuat Guneri, (2011) “ A weighted additive fuzzy programming approach for multi-criteria supplier selection”, Expert Systems with Applications, Vol. 38, pp. 6281–6286

[27] Zadeh, L.A., (1965), “Fuzzy Sets”, Information and Control. Vol.8, No.3, pp.199-249.

AUTHOR INDEX