In the Name of God

Journal of
Information Systems & Telecommunication
Vol. 5, No. 3, July-September 2017, Serial Number 19

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology

Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216

Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
E-mail: info@jist.ir
URL: www.jist.ir

Indexed by:

- SCOPUS www. Scopus.com
- Index Copernicus International www.indexcopernicus.com
- Islamic World Science Citation Center (ISC) www.isc.gov.ir
- Directory of open Access Journals www.Doaj.org
- Scientific Information Database (SID) www.sid.ir
- Regional Information Center for Science and Technology (RICeST) www.ricest.ac.ir
- Iranian Magazines Databases www.magiran.com

Publisher:
Regional Information Center for Science and Technology **(RICeST)**
Islamic World Science Citation Center **(ISC)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

Abdolvand, Neda, Alzahra University, Tehran, Iran

Baradaran, Vahid, Islamic Azad University, Science and Research Branch, Tehran, Iran

Charmi, Mostafa, Zanjan University, Zanjan, Iran

Ebadati, Omid, Kharazmi University, Tehran, Iran

Emadi, Mohammd Javad, Amirkabir University of Technology, Tehran, Iran

Farzi, Saeed, Khaje Nasir-edin Toosi University, Tehran, Iran

Hamidi, Hojjatollah, Khaje Nasir-edin Toosi University, Tehran, Iran

Kheirkhah, Fatemeh, ACECR, Tehran, Iran

Koosha, Hamidreza, Ferdowsi University of Mashhad, Mashhad, Iran

Momen, Amir, Islamic Azad University, Yadegar-Emam Branch, Shahre-Rey, Iran

Momtazi, Saeedeh, Amirkabir University of Technology, Tehran, Iran

Rafighi, Masoud, Malek-Ashtar University of Technology, Tehran, Iran

Rahmati, Mohammad, Amirkabir University of Technology, Tehran, Iran

Rashidi Kanan, Hamidreza, Shahid Rejaee Teacher Training University, Tehran, Iran

Rasi, Habib, Urmia University, Urmia, Iran

Rezvanian, Alireza, Institute for Research in Fundamental Sciences, Tehran, Iran

Sabbaghian, Maryam, University of Tehran, Tehran, Iran

Sadat Rasool, Seyed Mahdi, Kharazmi University, Tehran, Iran

Sadeghi Zavareh, Mostafa, Islamic Azad University, Najafabad Branch, Najafabad, Iran

Shayesteh, Mahrokh, Urmia University, Urmia, Iran

Shirvani Moghaddam, Shahriar, Shahid Rajaee Teacher Training University, Tehran, Iran

Shirzadian Gilan, Maryam, Islamic Azad University, Kermanshah Branch, Kermanshah, Iran

Speily, Omid, Urmia University of Technology, Urmia, Iran

Tavakoli, Hassan, University of Guilan, Guilan, Iran

Vahidipoor, Seyed Mahdi, University of Kashan, Kashan, Iran

Valizadeh, Mohammadreza, Ilam University, Ilam, Iran

Table of Contents

Representing a Content-based link Prediction Algorithm in Scientific Social Networks

Hosna Solaimannezhad*
Faculty of Electrical and Computer Engineering,University of Tehran, Tehran, Iran
hosna.sn91@yahoo.com
Omid Fatemi
Faculty of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
omid@fatemi.net

Received: 23/Apr/2017 Revised: 05/Sep/2017 Accepted: 08/Oct/2017

Abstract

Predicting collaboration between two authors, using their research interests, is one of the important issues that could improve the group researches. One type of social networks is the co-authorship network that is one of the most widely used data sets for studying. As a part of recent improvements of research, far much attention is devoted to the computational analysis of these social networks. The dynamics of these networks makes them challenging to study. Link prediction is one of the main problems in social networks analysis. If we represent a social network with a graph, link prediction means predicting edges that will be created between nodes in the future. The output of link prediction algorithms is using in the various areas such as recommender systems. Also, collaboration prediction between two authors using their research interests is one of the issues that improve group researches. There are few studies on link prediction that use content published by nodes for predicting collaboration between them. In this study, a new link prediction algorithm is developed based on the people interests. By extracting fields that authors have worked on them via analyzing papers published by them, this algorithm predicts their communication in future. The results of tests on SID dataset as co-author dataset show that developed algorithm outperforms all the structure-based link prediction algorithms. Finally, the reasons of algorithm's efficiency are analyzed and presented.

Keywords: Link prediction; Social networks; Content-based; Interest.

1. Introduction

As a part of recent study progress, a great deal of attention has been devoted to computational analysis of social networks. Indeed, a social network is a social structure composed of a set of social actors and set of communications between these actors. A social network can be imagined as a graph in which the nodes show the entities that are in a social context and the edges refer to the communication, interaction, collaboration or the effect between these entities. Any unit that could connect to other units could be considered as a node in the social network. One of the social networks is co-author network. In these networks, the nodes represent the papers' authors and the edges show the collaboration of these authors in writing papers. Like other social networks, these social networks are dynamic and they are changing over the time via adding the nodes and edges. The dynamics of these networks has turned them into a challenging topic for study. The network becomes even more complex as network nodes and edges grow, and they can be analyzed using the network analysis. In the social network analysis, analyzing the communication between this network's actors and analyzing the content published by these actors are the main concerns. Indeed, a social network analyst seeks to discover how their entities are created and connected to the social network. The social network analysts believe that the success or failure of a community depends on the structural patterns in the social network graph [1]. SNA fields are divided into descriptive and predictive, and they can focus on the links or social networks' entities. Link prediction is a prediction issue focusing on the links. Link prediction is a sub-branch of social network analysis being used in other fields as recommender systems, molecular biology and criminal researches. This is the only sub-branch of social network analysis focusing on the links instead of focusing on the entities. This makes the link prediction attractive and makes it distinguished from other data mining domains. Link prediction is a sub-set of link mining [2]. Link mining is a subset of data mining. Our field of study is to predict the network of collaborator writers. In these networks, the authors represent the network nodes and their collaboration is shown as link in the network graph. These networks show the collaboration between the papers' authors.

There are few studies on link prediction using the content published by nodes to predict the link. For this reason we have developed an algorithm based on the content published by nodes. To define the problem, suppose a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? We consider this question as the link prediction problem. This issue focuses on the links

* Corresponding Author

between the entities in network. Indeed, the collaboration prediction between the entities in the network is called link prediction. The goal of link prediction is to estimate the probability of creating links between the nodes in social networks [3], [2]. This estimation could be performed via analyzing attributes of entities and network structure. The meaning of network structure is the structure of other links in the network. In fact, the structural characteristics of the social network are the same as the topological properties of the network graph. In link prediction, the target is predicting links at time t_2, while we have links at time t_1 [4]. Any link prediction algorithm gives a score to a non-existing link and this score shows the probability of existence of the link at time t_2 and it calculates the similarity between the nodes that are in the end of the link. All the nonexistent links will be sorted in a descending order according to their scores, and the links at the top of the list are most likely to exist in the future [5], [2]. Considering a social network G (V, E) at time t1, where V is the set of nodes and E is the set of edges. The interaction between nodes u ,v is shown as edge e, as e ϵ E. Link prediction is defined as: The link prediction algorithm only by accessing the graph of the time interval t_1, should predict existence of the edges that exist in time interval t_2, but they haven't existed in time interval t_1. As $t_2>t_1$, set t1 is called training set and set t_2 is called test set.

In this study, different methods of link prediction have been analyzed and investigated and a content-based method has been presented for link prediction in co-authorship network.

2. Literature Review

From a specific view, link prediction methods are divided into four categories [6]: Node-based methods, topology-based methods, social theory-based methods and learning-based methods.

2.1 Node-based Methods

The calculation of similarity between a node pair is a solution for link prediction. This solution is based on a simple idea: nodes that are more similar to each other are more likely to have a link between each other. Indeed, people tend to create relationship with people who are similar in educations, religions, interests and locations. In this method, the similarity between non-connected pair of nodes in a social network is computed. This method is based on the criterion to analyze the proximity of nodes. Each pair (x,y) has a score and higher score means x , y are more likely to communicate with each other in the future,while lower score means that the nodes are more likely to have no link in the future. Thus, a list of scores in descending order will be achieved and the links at the top are most likely to exist in future. By this list, we can predict the links that will be created in future [6].

In a social network, a node has some attributes such as a profile in online social networks, mail name in e-mail networks and a series of published articles in scientific social networks. The information is used directly to calculate the similarity of two nodes. Since in most cases the values of node's properties are textual, typically, text-based and string-based similarity metrics are used. Papers [7], [8] have discussed about these criteria in details.

The authors [9] have defined a tree model to study the keywords of the user profile. They have used the distance between the keywords to estimate the similarity between the node pairs. They also have shown that by increasing the number of friends and keywords, the average similarity between the user and his friends decreases.

The authors [10] have found that most user profiles in current social networks are missed. To overcome this limitation, they have proposed a method, using stronger profiles, to infer some of the lost values, before calculating similarity.

The authors [11] have used overlapping user interests to measure the similarity. User interests are inferred from the actions they take, such as editing an article in Wikipedia. All actions that a user does can be shown as a vector, then, the similarity between two users will be obtained via the cosines similarity of their vectors.

Generally, node-based metrics use the attributes and actions reflecting the user interests to calculate the similarity between node pairs. These methods are useful if we can access to the user profile information and his performance, or we can infer them.

2.2 Topology-based Metrics

Even in networks where no information is available of nodes and edges, we can calculate the similarity between nodes by many other criteria, because the majority of criteria are based on graph topology and they don't need to know the attributes of nodes and edges. The graph structural attributes are defined in details in [12]. These methods are called similarity-based criteria. These methods use simple algorithms that, at the worst state, have time complexity of O (n^3). According to the attributes of these criteria, we can divide them into three groups of neighbor –based criteria, path-based criteria and random walk-based criteria [6].

In social networks, people tend to create new relationships with people that are closer to them. It is clear that the neighbors are the closest people to a social network user. For this reason, many neighbor-based criteria are developed by researchers for predicting links that will exist in future. For example, an algorithm called common neighbors is developed by the authors of paper [13]. In this algorithm, to estimate the similarity between the nodes, their common neighbors are computed. Other neighbor- based algorithms are generalizations of this algorithm. For example, the authors of paper [14] have introduced Adamic/Adar criterion to compute the similarity between websites. Paper [4] shows that Adamic/Adar (AA) is one of the best link prediction methods. After the extraction of the attributes of web-

pages, they will give the higher weight to the common attributes of two websites that is rare. It means that the attribute that is common between just two websites takes the highest weight. This criterion can be generalized to common neighbors of two nodes. In this way, the common neighbors between the two nodes, which are rare and shared only between the two nodes, take higher weight. Indeed, the common neighbors between the node pair that has lower degree take higher weight. Whenever the network studied is a co-author network, if an author has many co-workers, the probability of being shared between the nodes will be greater, but this probability will be very low for the nodes with lower degree.

This criterion is presented as [14]:

$$AA(x,y) = \sum_{z\in\Gamma(x),\Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (1)$$

Paper [4] shows that Adamic/Adar (AA) is one of the best link prediction methods.

In addition to node-based criteria and neighbor –based criteria, we can use the path between two nodes to estimate similarity between node pair. For example, authors of paper [15] have established Katz based on the influence of all paths. This criterion [15] counts all paths between all pairs of nodes. In this criterion, we can give more weight to shorter path, because it is obvious that: the longer the length of the path, the less impact will have in linking the nodes. The formula of this criterion is represented as [15]:

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l . |path_{x,y}^l| = \beta A_1 + \beta^2 A_2 + \beta^3 A_3 + \cdots \quad (2)$$

In this formula, $Path_{x,y}^l$ is the set of all paths from x to y that have length 1 and $\beta > 0$. If β is very small, this criterion will act similar to common neighbors, because long paths are not included in the calculation.

There are other criteria estimating the similarity between nodes via paths between them. Papers [16], [17], [18] have developed path-based criteria for link prediction.

The social relations between the social network nodes can be modeled using random walk. There are some methods computing the similarity between nodes in social networks based on random walk. These methods by defining a special destination for a random walk, from a special node, use the probability of going to the neighbors for prediction. For example, hit time (HT) is expressed by [19]. In this algorithm, the similarity between x,y nodes is estimated using calculation of the required walk number for a random walker, from the node x to node y. The smaller this number means the two nodes are more similar to each other. This method is formulated as [19]:

$$HT(x,y) = 1 + \sum_{w\in\Gamma(x)} P_{x,w}\, HT(w,y) \quad (3)$$

In this equation, Pi,j is the probability of going from node i to node j. Matrix P is defined as: $P=D_A^{-1}A$. In this formula, D_A is the diagonal matrix A in which $(D_A)_{i,i} = \sum_j A_{i,j}$. Clearly, the smaller this value is, the more similar the two nodes. So, in order to obtain the similarity between nodes, we multiply the value in negative.

The other random walk-based methods are expressed in papers [20], [21], [22], [23].

2.3 Social Theory-based Criteria

Social theory-based criteria can improve the efficiency of link prediction using additional information about social relationships. These methods are particularly suitable for large-scale social networks. In recent years, many researchers have applied old social theories, such as community, triadic closure, strong and weak ties, homeomorphisms, and structural balance, for analyzing and exploring social networks.

In paper [24], by considering user interest and user behavior, topology information is combined with community information. In this study, tweeter dataset is used for link prediction. They have shown that this method could improve the link prediction efficiency in directional, big scale networks.

In a paper [25], a link prediction model has been developed based on weak ties. It also uses three characteristics of the centrality of common friends, such as centrality, proximity, and betweenness. Each common neighbor, depends on their centrality, plays a different role in probability of communication between nodes. The weak tie is also considered for improving prediction accuracy. This model can be defined as follows [25]:

$$LCW(x,y) = \sum_{z} \big(w(z).f(z)\big)^{\beta} \quad (4)$$

F(z) is the switch function, and if z is common neighbor of x, y nodes, its value will be 1, otherwise its value will be zero. W(z) defines the centrality value of a node. β Parameter can moderate the quota of each common neighbor in probability of connecting two nodes. It is obvious that when this parameter is greater than 1, larger centrality values will be much more effective than smaller centrality values. When this parameter is less than 0, it restrains and prevents impacts of larger centrality values more than lower centrality values.

There are other social-theory based methods for link prediction. The authors of papers [26], [27], [28] have proposed some social theory-based criteria for link prediction.

2.4 Learning-based Methods

In recent years, many methods are presented based on learning. These methods use the external information and the attributes provided by algorithms considered in the previous sections for link prediction. These methods also create a training mechanism for prediction and consider special patterns that are special for each graph, in prediction. These algorithms have better efficiency compared to the previous algorithms, but due to the time-consuming training phase, they have high time complexity and sometimes they couldn't be applied on large networks.

For example, in paper [29], Support vector machine (SVM) is used for link prediction. The performance of this algorithm is in this way that each sample of paired nodes taken will be mapped to a point in the space. These samples have positive or negative label. So, two classes with empty gap are in the space. Now, the samples entering, based on their closeness to the classes, will be mapped to a class. This algorithm plots some hyper planes and attempts to extract a hyper plane showing the distinction between the classes better.

Generally, these methods are performed using a link prediction training mechanism.

3. Problem Solving Method

In this section, we explain the proposed method to extract people interest and method of interest vector formation for each author.

Generally, this section consists of six parts. First section is network preparation, as the algorithm can be applied to it. Second section is language processing of the network's content. Third section is extraction of papers' keywords. Fourth stage is extraction of the papers' topics and forming subject vector for papers. In the fifth stage, authors' study fields are extracted. Finally, in the sixth stage, based on the fields of authors, link prediction is performed. In the following, each of these sections is explained separately.

3.1 The Preparation Method of Network

The network that is used for this study is co-authorship network of SID site. This network consists of raw data in XML format. For each year, there is a XML file. At first, these files are integrated with each other, then to use this network, it is required to extract the graph of training time interval and testing time interval. The graph of training set includes papers' authors and their links during 2000-2005. The testing graph includes papers' authors and their links during 2006-2012. In this stage, different types of mapping are performed. These mappings include:

The mapping of papers to corresponding XML file: since this network is stored in XML format, using this mapping, the location of the article can be obtained in the corresponding XML file. By this mapping, we can easily get other information about the articles, such as keywords, relevant organization, the link to PDF format of the paper and gathering time.

The mapping of paper to the authors: Using this mapping, we can achieve the authors of each paper based on their ID.

The mapping of authors to the papers: Using this mapping, we can achieve the papers written by each author.

The mapping of papers to the papers' summary: Using this mapping, we can achieve the abstract of each paper.

Mapping papers to the keywords of papers: Using this mapping, we can achieve the keywords of each paper's abstract.

The mapping of papers to the root words of papers' keywords: Using this mapping, we can achieve root of the keywords in papers' abstract.

For these mappings, we use some Hash Map functions. By defining key and value for these functions, mappings are performed.

3.2 Language Processing of the Network Content

In this section, the abstract of each paper is processed by language processing to be used in the next steps for extracting articles from articles. In this stage, Persian processing tool, developed by [30] in telecommunication research center, is used. This system can do all the necessary actions for different layers of Persian language processing, from its initial layer which is lexical layer up to the upmost layer which is syntax. This Toolkit performs a combination of these processes: normalization, tokenization, Spell checker, morphological analysis, Persian Dependency Parser, Semantic Role Labeling. This toolkit, by receiving the Persian raw data, performs semantic and morphological analyses. These analyses include normalization, Tokenization, stemming and Lemmatizing. Thereupon, by adding this information to raw text, by Dependency Parser ParsiPardaz, Dependency Parse Tree is generated for Persian sentences.

The processes we've been using with this toolkit include the following:

Text normalization: The characters of words that are in the abstract are normalized in this stage. For example, converting Arabic ی to Persian ی or converting the characters "إ"، "أ"، "آ" to a unified form "ا". As the same way, we convert the words "سِر" و "سُر"، "سَر" to a uniform word "سر". Before any language processing, the upper level of this conversion should be performed. It means that the similar characters should be unified. This challenge that exists in Persian language is called Unicode Ambiguity. To solve this problem, Lemmatizing is used to perform normalization task.

Tokenization: In this stage, text sections are defined. For this stage, Persian language processing toolkit, developed by [30] in telecommunication research center, is used. This toolkit performs text tokenization based on syntactic dependency rules and some semantic and syntactic features. In this method, all compound words are connected by hemi-space. For example, instead of" آمده است", آمده‌است" is used. Since in the next steps, the extraction of subjects, the border between words is specified by a Space character, thus, tokenization is used which specifies the boundary of words by putting together compound words with hemi-space and putting the Space character, the border of words is defined.

These two tools are in the first level of Persian Language processing toolkit ParsiPardaz, lexical layer.

3.3 Keyword Extraction

In this stage, Stop words, the words that are common and not useful, will be eliminated from the abstract. At

first, all punctuation marks are eliminated from the text. Then, stop words will be deleted from the abstract's words. To extract stop words, the different steps have been taken. These steps include:

Eliminating common Persian words: There is a list of common Persian words that using this list consisting 500 words, a part of common words has been eliminated from the abstract [31].

Eliminating some respect words such as Engineer, Doctor, etc.

Eliminating some verbs gaining a list of common words among all papers using tf/idf method: This step, depending on the textual content of each network, can provide a different list of words to the user.

Extracting root of keywords: in the next steps, keywords of abstract are used in two ways: Once by considering keywords without the root of words and another time by considering their root. we can use root of the words instead of the words, as keywords. In this way, some words like subject and subjects are considered to be the same and it will be effective in calculating similarity between vectors of people's study field. To extract the root of words, stemming, the developed toolkit in telecommunication research center is used [30]. Stemmer tool is located in the second level of this toolkit, Morphology Layer. This tool applies the word structure to extract the root of words and acts independent from its content.

This step is one of the important sections in our algorithm.

3.4 Extracting Articles' Topic

In this section, we label all existing articles in dataset with the extracted topics. Indeed, for each paper, its interference with the topics is calculated. The set of documents is denoted by D and the set of words in the domain is denoted by V. In this step, this set includes the extracted keywords in the previous step, because the input words of the LDA algorithm are more useful and show the concept of the document more effectively, the algorithm will perform better. A denotes the set of authors and H denotes the set of extracted topics in this section. Each author participates in writing some of papers that are member of the D set. Each author that is a member of A set collaborates in writing the set of papers and this set is denoted by D_a and a denotes the code of author of this set of papers. To extract the subject of papers the developed LDA algorithm in paper [32] is used. LDA is used due to its superiority compared to the similar methods of topic extraction. This topic model is used on discrete sets such as textual sets. LDA is applied on many topics such as Collaborative Filtering, Text Classification, Word Sense Disambiguation [33] and Community Recommendation [34]. LDA is a Generative Model for the text and other Discrete Data Collections. In the context of textual modeling, this model claims that each document is produced in a combination of subjects. So, this algorithm returns interference level between documents and subjects as output by taking the document text and the number of requested subjects.

This algorithm defines a matrix $|V| * K$ for each paper. The elements of this matrix are consistent with the initial values of relevance between the words and topics. These values are considered similar for all subjects. The optimal value of this parameter will be investigated in the next chapter. This algorithm defines a matrix for each word with dimensions $|D| * K$. The elements of this matrix are consistent with the initial values of relevance between documents and topics. These values are considered similar for all topics. The optimal value for this parameter will be investigated in the next chapter.

LDA considers each document as polynomial distribution on topics. For each word existing in the document, it gives the topic distribution on the words, it shows a matrix in which rows are words and columns are topics. K is one of the inputs of algorithm. It shows the number of topics that we expect the algorithm to extract. If we define k topics, a matrix $|V| * K$ is defined showing the distribution of topics on the existing words in the domain of words. This matrix is denoted by φ. The probability of existence of the word w on topic k_{th} is denoted by $\varphi_{w,k}$. This probability values are achieved by applying LDA algorithm to the initial matrix, several times. The optimal value of the number of these iterations will be investigated in the next chapter. Then, using this distribution, the distribution of document on the topics is computed using the used words in this document and distribution of topics on these words. Indeed, a matrix with dimensions $|D| * K$ is defined. The rows of this matrix are the papers and the columns are topics. Each element of this matrix indicates the probability of belonging an article to a special topic. This matrix is denoted by θ. The probability that document d_i is related to k_{th} topic is denoted by $\theta_{i,k}$. For each paper the sum of these values is 1. The values are achieved by applying LDA algorithm on the initial matrix, several times. The optimal value of these iterations will be investigated in the next chapter. In this study, we apply LDA to the extracted keywords of the existing papers in the network to have a set of users' interests and to increase the accuracy of link prediction algorithm. Generally, the fewer the number of defined topics is, in each topic there are more concepts, and the topics will therefore be more general and the more the number of topics is defined, the less the concepts within these topics will be and the topics will be more specific. The effect of defined topics will be evaluated in the next chapter.

Generally, for each social network user and for each paper, we define a vector denoting the topic distribution of the paper. Then, topic distributions are used to achieve the study fields of authors and then we use the similarity estimation of authors' study field for link prediction. This topic distribution is shown as a matrix that its elements show the thematic interference of papers with the defined topics. Finally, as an output of this step of the algorithm, we have the subject vector of the same number of articles

existing in the data set and the elements of these vectors show the interference of document with extracted topics. These vectors are extracted from matrix θ achieved by LDA algorithm. Indeed, each row of this matrix represents the corresponding vector of a paper.

For example, the topic vector is defined for d_i paper as:

$$S_i = [S_{i,1}, S_{i,2}, \dots, S_{i,|H|}]$$

These values show the similarity between document d_i and different topics. These values are between zero and one and define the relevance of document d_i to different study fields. Also, sum of the values of these elements is one.

3.5 Extracting Authors' Study Fields

In this step, using the vectors extracted in the previous step, for all authors, we make an interest vector. The elements of these vectors show the interest of the given author in the given study field. The number of elements in these vectors is equal with the number of elements in subject vectors of the papers. Indeed, using the topics of each paper, the study field of each author could be defined. Accordingly, the study field of each author is derived from the average of the subjects of all articles written by him. To extract the study field of authors, various methods, such as maximizing between the subjects of articles written by author, have been evaluated and this method has been the best.

It should be noted that the number of extracted topics for all articles and authors is the same and the topics are similar. For example, if 100 topics are defined for each paper, all authors have these 100 topics. The relevance of each author to each topic is computed using the probability number assigned to that subject for articles written by that author. Each person is the author of some papers. To achieve the activity of authors in different study fields, we compute the mean of the corresponding elements in the subject vector of the articles published by this author and we put them in a vector called Interest Vector. The interest vector is defined for the author a_i as:

$$I_i = [I_{i,1}, I_{i,2}, \dots, I_{i,|H|}]$$

These values show the belonging of a_i author to different topics.

The elements of vector I_i are computed as follows: For example, the first element of a_i authors' interest vector, if this author participates in writing the papers d1, d2, d3, d4, it is computed as:

$$I_{i,1} = \text{AVG}\,(S_{1,1}, S_{2,1}, S_{3,1}, S_{4,1})$$

Finally, this step's output is the amount of the authors' activity on extracted topics, as interest vectors for the authors. In the last section of this chapter, using these vectors, the level of similarity between the authors is estimated to predict the collaboration between them.

3.6 Link Prediction Based on the Similarity of Authors' Study Fields

In this method, the probability of creating link between two authors who have not previously collaborated is consistent with the similarity between study field vectors of these two authors. To calculate the similarity, Cosine Similarity formula is used. Other methods as Euclidean similarity have been evaluated and finally, Cosine Similarity has generated the best results. Cosine Similarity is a similarity criterion between two vectors which calculates the cosine of the angle between two vectors. Zero's cosine is equal to one, thus if two vectors are match in each other, their similarity is equal to one. It is clear that this value shows the highest similarity between two vectors. Indeed, if the interest vectors of two authors are consistent, the similarity of these two authors is considered as 1. For any other angle, this value is less than 1. If two vectors with angle 90 degree are in the space, their cosine similarity is zero. It is obvious that this value shows the lowest similarity between the vectors. Since cosine similarity is computed in positive space, the similarity between vectors will be between zero and one. In information retrieval and Text Mining, this criterion is used to estimate the similarity between document vector and query vector or the similarity between the vectors of two documents [35]. Also, in data mining, this method is used to estimate the coherence of clusters [36]. The reason to use cosine similarity is that this criterion is effective on evaluation, especially for sparse vectors, because only non-zero values are considered. Since the interest vectors of authors are sparse, we use this criterion. Cosine similarity of two vectors is computed as [35]:

$$\text{CosineSim} = \frac{\sum_{i=1}^{K} A_i * B_i}{\sqrt{\sum_{i=1}^{K} (A_i)^2} * \sqrt{\sum_{i=1}^{K} (B_i)^2}} \quad (5)$$

In this formula, A, B are study field vectors of two authors. K is the number of topics.

Based on the following formula, we compute the similarity between two authors with each other:

$$\text{Score}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\text{cosine}(x, y)}{\log(|\Gamma(z)|)} \quad (6)$$

According to this formula, if there is much similarity between the topics that the two authors are interested in and the degree of the common node between two authors is low, they will be more similar to each other and it is highly probable that they will establish a link in future. The reason is that the lower the degree of the common author, the better it is and it shows that the mentioned node has higher similarity to these two authors, rather than the case in which common author has many common authorships. This criterion is the combination of content similarity between the nodes with Adamic- Adar criterion. Indeed, in Adamic- Adar criterion, the content similarity of the authors is not involved in link prediction.

This method is called Structure Topics Prediction.

4. Experiments and Results

We used a useful dataset in order to applying our algorithm to this. The students of DBRG Lab, a Lab in Tehran University, had produced a dataset that is extracted from co-authorships between authors of existing papers in Academic Jihad Scientific Information Database. The developers of this dataset have named it as SID. The generated graph of this dataset is not weighted and not directed. This dataset is related to articles from 1379 to 1390. In order to link prediction we divided the data into two periods of training and testing. These two periods are as follows: collaborations between the authors from 1379 to 1374 - collaborations between the authors from 1385 to 1390. The output of a link prediction algorithm is a sorted list of scores allocated to links not existing in the training time graph. One of the methods being used to evaluate the results of link prediction algorithms is the area under Receiver Operating Characteristic (ROC) curve [37]. This method is called AUC (Area under Receiver Operating Characteristic). The horizontal axle of ROC chart shows the false positive links and in the vertical axle, true positive links are considered. The false positive links are called FP and the number of true positive links is called TP. In this method, an algorithm has a better output that gives higher score to the links created in the testing time interval than the links not created in this time interval. The number of links not existing in training set is denoted by n. Also, n_1 is the set of the links created in the test set and n_2 is the set of links not created in the test set. We achieve all pairs $n_2 * n_1$ and represent the number of them with k. If in m number of k pairs, the score given to the existing link is higher than the score given to the non-existing link and in b numbers, it is opposite. AUC criterion is achieved as [38]:

$$\text{AUC} = \frac{\left(m + \frac{b}{2}\right)}{k} \tag{7}$$

Another method for evaluating link prediction algorithms is Area under Precision-Recall Curve (AUPR). This criterion uses the area under Precision-Recall chart. The precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + TN} \tag{9}$$

In fact, precision is equal to the percentage of predicted links that are correctly predicted and recall is is equal to the percentage of the links generated in the test interval that are predicted. Indeed, in recall points, the precision of the algorithm is measured and the chart is plotted.

In the following, we compare the proposed algorithm with other algorithms in the domain of the link prediction using the criteria mentioned.

Based on the results achieved from previous sections, we compare the results of proposed algorithm with other structural link prediction algorithms. In the following Table, for Katz algorithm, parameter β is 0.005 and the maximum distance from the source node is considered as 5. For RootedPageRank, the probability of return to the previous node is considered as 0.5. For StructureTopicsPrediction algorithm, α is 0.5 and β is 0.001, the number of topics is 150 and the number of iterations of LDA algorithm is 200.

Table 1. the results of proposed algorithms with other algorithms

Algorithm	AUPR	AUC
StructureTopicsPrediction	0.033441885	0.735185072
Common neighbors	0.028207159	0.630440334
Adamicadar	0.034373253	0.712968505
Jaccard Coefficient	0.009238340	0.467782634
Distance	0.010419157	0.500000000
PreferentialAttachment	0.017296527	0.627418673
Katz (5-0.005)	0.029558192	0.648661437
RootedPageRank (0.5)	0.014508203	0.612446540

As shown, Adamic-Adar and Katz algorithms have good results. In the comparison of the proposed methods with other methods, we can achieve the following results:

The combination of content and structure can improve prediction outcomes. The results of proposed algorithm indicate this matter.

Our method can improve Adamic- Adar, as the best algorithm between the other algorithm, results about 3% and can improve Katz,s results about 10%. The combination of content and structure can improve the prediction results. The results of Structure Topics Prediction algorithm are good examples.

The best AUC criterion belongs to Structure Topics Prediction. According to the studies [37], the importance of AUC criterion is higher than the importance of AUPR and the algorithm with the better AUC has good results. Although AUPR criterion shows the area under precision-recall chart, but AUPR chart is effective on the comparison between the performances of algorithms. Because it is possible that an evaluation method has the best value of AUPR, but in some cases it has lower quality than the other algorithms and vice versa. For this purpose, the following diagrams are used to evaluate the results of the algorithms more precisely. AUC, P-R charts are plotted to compare the proposed algorithm with two algorithms having the best results.

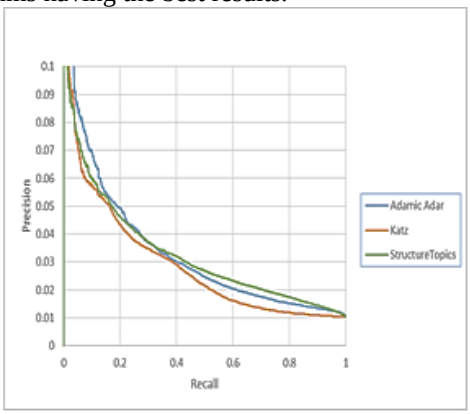


Fig. 1. ROC Chart

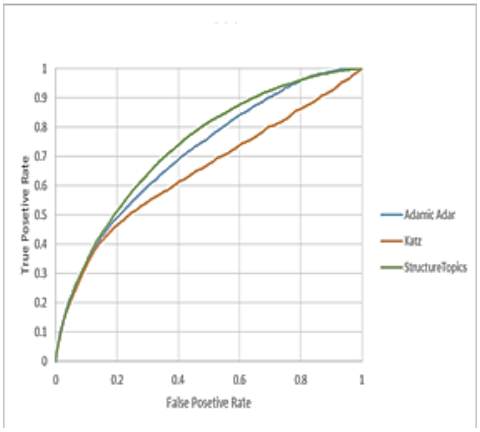


Fig. 2. P-R Chart

According to the results of these evaluations and the optimal adjustment of the input parameters of LDA algorithm, this method is compared with the existing methods in link prediction. The results of this comparison show that the combination of content and structure can increase accuracy of link prediction algorithms.

5. Conclusion and Further Studies

Based on the purpose of this study, the achievements of this study are: Different methods of link prediction have been studied. A link prediction algorithm based on the content and interests of people has been presented. The comparison between the performance of existing algorithms and presented algorithm in this study has been evaluated through several evaluation methods and finally, the outputs are analyzed.

In future studies, to improve the quality of algorithms, we can work effectively on extracting the subject of the content published by the network nodes and recover the network content effectively. For example, we can extract the keywords of texts exactly and to improve the content-based algorithm results, we can improve the extraction of keywords. Also, we can use the data fusion algorithms to combine the results of different algorithms and achieve better results. Also, the data fusion algorithms can be used in allocating study fields to the authors. Generally, to improve this section of algorithm, the extraction of authors' study fields from the topic of their study, we can make efforts.

It is possible that a link predicted by a method is created in a period after test set. Indeed, the link is predicted true but it is not created at appropriate time. The current methods for evaluating the accuracy of algorithms did not consider this matter. Considering this case, to some extent, could provide a more accurate assessment of the accuracy of the algorithms.

References

[1] F. Borko (Ed.), Handbook of Social Network Technologies and Applications, 2010 .
[2] L. Getoor, Christopher P. Diehl“ , Link Mining: A Survey ”, SIGKDD Explorations ,Vol.7 ,No. 2 ,pp. 3-12, 2005 .
[3] B.Taskar, MF Wong, P Abbeel, D Koller“ , Link prediction in relational data ”, Learning Statistical Patterns in Relational Data Using Probabilistic Relational Models , Vol.7 ,2005 .
[4] D. Liben-Nowell, J. Kleinberg“ ,The Link Prediction Problem for Social Networks ”,Journal of the American Society for Information Science and Technology , Vo.No.58,7 ,pp. 1019-1031, 2007 .
[5] X. Feng, Z.J., Xu K“ ,Link prediction in complex networks: a clustering perspective ”,European Physical Journal , vol.85 ,pp. 1-9, 2012 .
[6] P. Wang, B. Xu, Y. Wu, X. Zhou“ ,Link Prediction in Social Networks: the State-of-the-Art ”,Science China Information Sciences ,Vol.57 ,pp. 1-38, 2014 .
[7] V. StröEle, G. ZimbrãO, J. Souza“ ,Group and link analysis of multi-relational scientific Social Networks ”,Journal of Systems and Software ,Vol.86 ,pp. 1819-1830, 2013 .
[8] G. Rossetti, M. Berlingerio, F. Giannotti“ ,Scalable link prediction on multidimensional networks ”, in11th IEEE International Conference on Data Mining Workshops , Vancouver, Canada, 2011 .
[9] J. Mori, Y. Kajikawa, H. Kashima, “Machine learning approach for finding business partners and buildings reciprocal relationships ”,Expert Systems with Applications ,Vol.39 ,pp. 10402-10407, 2012 .
[10] W. Sen, J. Sun, J. Tang“ ,Patent partner recommendation in enterprise social networks ”,inthe 6th ACM International Conference on Web Search and Data Mining (WSDM'13) , Rome, Italy, 2013 .
[11] L. Aiello, A. Barrat, R. Schifanella,“ ,Friendship prediction and homophily in social media ”,inACM Transactions on the Web , 2012 .
[12] H. Chen , D.Miller, C.Giles“ ,The predictive value of young and old links in a social network ”, inthe ACM SIGMOD Workshop on Databases and Social Networks , New York, USA, 2013 .
[13] D. Davis, R. Lichtenwalter, N. Chawla“ ,Supervised methods for multi-relational link prediction ”,Social Networks Analysis and Mining ,Vol.3 ,pp. 127-141, 2013 .
[14] L. Adamic, and E.Adar “ ,Friend and Neighbors on the Web ”,Social Networks ,Vol.25 ,pp. 211-230, 2003 .
[15] P. Soares, R. Prudêncio“ ,Proximity measures for link prediction based on temporal events ”,Expert Systems with Applications ,Vol.40 ,pp. 6652-6660, 2013 .
[16] E. Richard, N. Baskiotis, T. Evgeniou,“ ,Link discovery using graph feature tracking ”,inthe 24th Annual Conference on Neural Information Processing Systems 2010 ,Vancouver, Canada, 2010 .
[17] S. Oyama, K. Hayashi, H. Kashima“ ,Cross-temporal link prediction ”,inthe 11th IEEE International Conference on Data Mining (ICDM'11 ,(Vancouver, Canada, 2011 .
[18] P. Ricardo ; Ricardo Bastos Cavalcante Prudêncio, “ ,Time series based link prediction ”, in International Joint

Conference on Neural Networks (IJCNN'12 ,(Brisbane, Australia, 2012 .

[19] E. Gilbert, K.Karahalios“ ,Predicting tie strength with social media ”, inthe SIGCHI Conference on Human Factors in Computing Systems ,Boston, USA, 2009 .

[20] J. O'Madadhain, J. Hutchins, P. Smyth“ ,Prediction and ranking algorithms for event-based network data ”,ACM SIGKDD Explorations ,Vol.7 ,pp. 23-30, 2005 .

[21] S. Brin , L. Page“ ,The Anatomy of a Large-Scale Hypertextual Web Search Engine ”,Computer Networks and ISDN Systems ,Vol.30 ,pp. 107-117, 1998 .

[22] R. Lichtenwalter, J. Lussier , N. Chawla“ ,New perspectives and methods in link prediction ”,in16th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining ,Washington, DC, USA, 2010 .

[23] D. Dunlavy, T. Kolda, E. Acar“ ,Temporal link prediction using matrix and tensor factorizations ”,ACM Transactions on Knowledge Discovery from Data , Vol.5 ,pp. 1-27, 2011 .

[24] T. Kuo, R. Yan, Y. Huang, “Unsupervised link prediction using aggregative statistics on heterogeneous social networks”, inthe 19th ACM SIGKDD international conference on Knowledge discovery and data mining , Chicago, USA, 2013 .

[25] D. Yin, L. Hong, B. Davison“ ,Structural link analysis and prediction in microblogs ”,inthe 20th ACM International Conference on Information and Knowledge Management (CIKM'11 ,(Glasgow, UK, 2011 .

[26] R. Xiang, J. Neville, M. Rogati“ ,Modeling relationship strength in online social networks ”,inthe 19th International Conference on World Wide Web (WWW'10) , Raleigh, USA, 2010 .

[27] “ Z. Yin, M. Gupta, T. Weninger, J. Han” LINKREC: a unified framework For link Recommendation with user attributes and Graph structure ”, inthe 19th International Conference on World Wide Web (WWW'10) ,Raleigh, USA, 2010 .

[28] M. Sachan, R. Ichise“ ,Using semantic information to improve link prediction results in network datasets ”, International Journal of Computer Theory and Engineering, Vol.3 ,pp. 71-76, 2011 .

[29] C. Cortes, V. Vapnik“ ,Support-vector networks ”,Machine learning ,Vol.3 ,pp. 273-297, 1995 .

[30] Z. Sarabi,H.Mahyar,M.Farhoodi“ ,ParsiPardaz: Persian Language Processing Toolkit ”,2013 .

[31] “ http://www.ranks.nl/stopwords/persian ”,[intra-linear].

[32] Xu. HieuPhan,L.Nguyen,S. Horiguchi .“ ,Learning to Classify Short and Sparse Text&Web with Hidden Topics from Large-scale Data Collections”, inThe 17th International World Wide Web Conference, Beijing, China, 2008 .

[33] J. Boyd-Graber, D. Blei, X Zhu“ ,.A Topic Model for Word Sense Disambiguation ”, inthe 2007Joint Conf. on Empirical Methods in Natural Language Processing and Comp. Natural Language Learning ,2007 .

[34] W. Chen, J. Chu., J. Luan, H. Bai, Y. Wang, Y.Chang ,. “Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior ”, inInternational World Wide Web Conference ,2009 .

[35] S. Amit“ ,Modern Information Retrieval: A Brief Overview ”,Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,, NO. 244 ,pp. 35-43, 2003 .

[36] P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, 2005 .

[37] D. Hand“ , Measuring classifier performance: a coherent alternative to the area under the ROC curve ”,Machine learning ,pp. 103-123, 2009 .

[38] J. Davis, M. Goadrich“ ,The relationship between Precision-Recall and ROC curves ”, inthe 23rd international conference on Machine Learning , 2006 .

Hosna Solaimannezhad received the B.Sc. degree in Information Technology engineering from University of Sepahan , Isfehan, Iran, in 2010. She received the M.Sc. degree in Information Technology engineering from Tehran University, Tehran, Iran, in 2015. She is currently Ph.D student in Department of Electrical and Computer Engineering, Tehran University. Her area research interests include Data Mining, Information Retrieval, Big Data, Distributed systems and Social Networks. Her email address is:

Omid Fatemi received the B.Sc. degree in Electrical engineering from Tehran University, Tehran, Iran, in 1989. He received the M.Sc. degree in Electrical engineering from Tehran University, Tehran, Iran, in 1991. He received the Ph.D degree in Electrical and Computer engineering from University of Ottawa, Ottawa, Ontario, Canada, in 1999. Now, he works as assistant professor in Department of Electrical and Computer Engineering, Tehran University. His area research interests include Big Data, IT Governance, E-Learning and cloud Computing.

A RFMV Model and Customer Segmentation Based on Variety of Products

Saman Qadaki Moghaddam
Department of Electrical, Computer and IT Engineering, Qazvin Azad University, Qazvin, Iran
sam13671367@yahoo.com
Neda Abdolvand*
Department of Social Science and Economics, Alzahra University, Tehran, Iran
n.abdolvand@alzahra.ac.ir
Saeedeh Rajaee Harandi
Department of Social Science and Economics, Alzahra University, Tehran, Iran
saeedeh.rh@gmail.com

Received: 19/Dec/2016 Revised: 18/Aug/2017 Accepted: 25/Aug/2017

Abstract

Today, increased competition between organizations has led them to seek a better understanding of customer behavior through innovative ways of storing and analyzing their information. Moreover, the emergence of new computing technologies has brought about major changes in the ability of organizations to collect, store and analyze macro-data. Therefore, over thousands of data can be stored for each customer. Hence, customer satisfaction is one of the most important organizational goals. Since all customers do not represent the same profitability to an organization, understanding and identifying the valuable customers has become the most important organizational challenge. Thus, understanding customers' behavioral variables and categorizing customers based on these characteristics could provide better insight that will help business owners and industries to adopt appropriate marketing strategies such as up-selling and cross-selling. The use of these strategies is based on a fundamental variable, variety of products. Diversity in individual consumption may lead to increased demand for variety of products; therefore, variety of products can be used, along with other behavioral variables, to better understand and categorize customers' behavior. Given the importance of the variety of products as one of the main parameters of assessing customer behavior, studying this factor in the field of business-to-business (B2B) communication represents a vital new approach. Hence, this study aims to cluster customers based on a developed RFM model, namely RFMV, by adding a variable of variety of products (V). Therefore, CRISP-DM and K-means algorithm was used for clustering. The results of the study indicated that the variable V, variety of products, is effective in calculating customers' value. Moreover, the results indicated the better customers clustering and valuation by using the RFMV model. As a whole, the results of modeling indicate that the variety of products along with other behavioral variables provide more accurate clustering than RFM model.

Keywords: Clustering; Data Mining; Customer Relationship Management; Product Variety; RFM Model.

1. Introduction

Today, increased competition between organizations has led them to seek a better understanding of customer behavior and their partners through innovative ways of storing and analyzing the customer information [1]. One of the major challenges of customer relationship management (CRM) is to establish more profitable and long-term relationships with customers [1] [2]. The emergence of new computing technologies has brought about major changes in organizations' ability to collect, store and analyze large datasets. Therefore, over thousands of data can be stored for each customer, enabling the analysis of customer purchasing history [3]. To understand and identify valuable customers, they should be segmented based on their behavior [4] [5]. The most frequently used model in customer segmentation is the RFM model, which consists of three behavioral variables: R (Recency), F (Frequency) and M (Monetary) [4]. Buying goods or services is the organization's customer communication channel. It can be used, along with other behavioral variables, to better understand and categorize customers' behavior, and adopt appropriate strategies such as up-selling and cross-selling. Since the RFM model is incomplete, studies have tried to improve and develop the model by adding other variables. Although, purchasing goods and services is the main channel of communication in businesses, few studies have considered the importance of identifying the products purchased. Since the variety of products is one of the main parameters of assessing customer behavior, studying this factor in the field of business-to-business (B2B) communication represents a vital new approach. Accordingly, based on the sales strategies of up-selling and cross-selling as the important tools of increasing customer value to organizations, this study tries to improve the RFM model by adding the variable of product variety (V). This will help customer recognition and provide more accurate understanding of the customer in the field of B2B.

* Corresponding Author

For the purpose of this research, the literature is reviewed first. The research methodology then addresses the model development and measurement tools. After the statistical analysis is explained, the conclusions and recommendations of the study are discussed.

2. Literature Review

In recent decades, social and economic changes in the interaction between organizations and customers have made customer relationship management (CRM), one of the most important processes in the business environment (followed by business partner relationship management) [3]. CRM can be defined as the collecting, storing, and analyzing customer data in order to increase customers' loyalty and value, and to increase organizational benefits [4] [6] [7]. In order to increase customer satisfaction and prevent customers from leaving the organization, the organization should focus on segmentation and meeting customers' individual needs [8]. Generally, "customer segmentation" is the process of dividing customers into different groups based on geographical, demographic and ethnological information, in order to adopt strategies tailored to each group based on the consumption of goods and services and customers' purchase history [4] [8] [9]. The segmentation process continues to achieve a harmony in customer value; along with detecting and correct placement of clients in related groups, it is one of the most essential factors in CRM and business success [4] [9]. Since there isn't a preferred approach to customer segmentation, the best model is one that draws a proper insight of current and potential customers, and help organizations to achieve effective markets and appropriate customer feedback [10]. Different techniques and forms of data analysis can be used for customer segmentation; the most common are the use of data mining techniques and customer behavior variables. Data mining is the process of discovering and extracting hidden patterns from large amounts of data [2]. The RFM model, which tries to identify customers based on their behavioral characteristics, is one of the most widely used [4] [11] [12]. This behavioral model uses three criteria of customer transactions: recency, frequency and monetary. These categories are then interpreted by data mining tools and algorithms. Thus, the RFM model predicts the customer's next movements based on their behavior [13]. It is not only considered as a segmentation model, but also includes the concept of customer value. That is why many studies use this model for the discovery and analysis of customer value based on past purchase behavior [13]. On the basis of their purchasing behavior, the RFM reveals that valuable customers are those with the highest frequency and monetary value and the lowest recency. Despite its simplicity of understanding, interpretation, and implementation [14] [15] [16], the model has shortcomings, such as inattention to personal characteristics and demographic variables of customers [14] [16]. Thus, studies have been done with the aim of increasing the accuracy of RFM model output by adding variables which can be categorized in two groups of customer-oriented and product-oriented. In the first group, factors such as customer lifetime value and imposed costs (eg, [4] [14] [17] [18] [19] [20] [21]), are taken into consideration, and in the second group, product life cycle (eg, [16]) as well as the types of products are considered as the basis for the development of model [14] [16]. Cheng and Chen [13] based their study on RFM model, and, by adding customer credit to the model, segmented customers of an e-services company in Taiwan. After customer clustering using their proposed LEM2 algorithm, they extracted important rules for future marketing decisions and company's strategies. Moreover, Khvajvand et al [1] in their study in the cosmetics industry have classified and valuated customers based on the number of products purchased by customers along with other behavioral variables in their proposed model. Given the RFM variables. Chang and Tsai [22] generalized RFM model to the groups of products and services, and proposed a concept, named Group RFM or the GRFM, to discover better customer consumption behavior. They believed that the final value of customers purchase should be calculated, and behavioral variables should be classified based on purchases. Noori [11] added the variable of customer deposit to the RFM model to classify mobile banking users. He indicated that identifying customers by a behavioral scoring facilitates marketing strategy assignment.

Most studies used customer-business approach (B2C) to develop model, and a few studies have been conducted in the field of business-business (B2B). In the field of B2B, the firm will reap huge profits from the large volume of purchases. Therefore, understanding customers would be the success key in his area. Some researchers proposed reasons such as quality of communication, trust, participation, satisfaction, increased buying, and organizational changes as the effective factors of maintaining customers in B2B organizations [16]. Hosseini et al. [16] have added product life cycle / product duration to RFM model, and classified customers into 34 categories in a B2B company. Kandeil et al. [4] also used LRFM clustering techniques to segment customers of a B2B distributor. Moreover, Venkatesan and Kumar [17] have studied the customer lifetime value of a B2B Manufacturer, and indicated that selecting customers based on the CLV provides more profits than other metrics. RFM evolution is briefly shown in Table 1.

Table 1. RFM Model Evolution

Added Variable	Resource	Industry
Customer Lifetime Value	(Ansari & Riasi, 2016)	Steel Company
	(Kandeil et al., 2014)	Distribution Company
	(Reinartz & Kumar, 2000)	Distribution Company
	(Alvandi et al., 2012)	Bank
	(Venkatesan & Kumar, 2004)	International Software and Hardware Producing Company
	(Wei et al., 2012)	Children's Dental Clinic
Customer Lifetime Value and Imposed Costs	(Soeini & Fathalizade, 2012)	Insurrance

Added Variable	Resource	Industry
Groups of Products and Services	(Chang and Tsai, 2011)	Educational Organization
Number of Products	(Khvajvand et al, 2011)	Cosmetics Industry
Customer Credit	(Cheng and Chen, 2009)	E-Services Company
Product Life Cycle / Product Duration	(Hosseini et al., 2010)	
Customer Deposit	(Noori, 2015)	Bank

According to studies carried out in this context, few studies have considered the importance of the product and their varieties, and reasonable valuation of customers based on product variety is not provided. In addition, most surveyed industries were in the field of B2C, and there has been insufficient attention to B2B domains. Therefore, with regard to the importance of the variety of products in leveraging strategies such as up-selling and cross-selling, increasing profitability of organizations, and customer maintenance, this study aims to add the variable of variety of products as a new behavioral variablein the RFM model, and to classify customers of the company in the field of B2B.

3. Methodology

Different algorithms and techniques are being used to classify, valuate, and model customer buying behaviors which are classified in two groups of clustering and association rule mining. Accordingly, clustering techniques and K-Means algorithms are used in this study. The CRISP-DM methodology, which is one of the greatest analytical methods for data mining projects, is used in this study. Moreover, K-Means algorithm and Silhouettes' measure of clustering quality is used to measure and to determine the number of clusters. This is done on 924 normalized records of customers by using SPSSModeler.14 software. Then, the ANOVA test is run on obtained clusters, and Duncan's post hoc test is used to determine distinct clusters using the SPSS.16 software. Finally, hierarchical analysis method is used for weighting the R, F, M, V variables, and to calculate the value of each customer.

3.1 Data Analysis

This study was conducted based on the CRISP-DM methodology which consists of six phases of business understanding; namely data understanding, preparation, modeling, evaluation and deployment [23] [24]:

Phase 1. Business Understanding

At this phase an overview of the type of business, Based on which the research is done is obtained, and the overall targeting is done based on the current strategies and business nature [24]. The objective of this study is to evaluate the variety of products purchased by customers along with other behavioral variables of the RFM model in the field of food and sanitary distribution company in the field of B2B to determine the similarity of customers based on their equity value, and to make company capable to identify high-value customers.

Phases 2 and 3. Data understanding and Preparing

The second phase involves collecting, describing and evaluating the data quality. In general, the aim of this phase is to select the appropriate data source in order to reach the goal [24] [25]. Thus, using the information available in our database, the data of over 1,000 customers, including customer code, name and total amount of purchase, are recorded in the database. In this stage, the output reports are received using Microsoft Excel 2010, and after the initial data sorting, the index values are extracted. Finally, the required variables, namely, the number of customers' purchases, customer's last purchase date, purchase amount, and the variety of purchased products in the last 6 months are extracted. Then data are prepared for modeling (third phase). Data preparation includes the process of excluding outliers and data normalization. In this study, 1112 records are collected; the number of data after removing duplicates data is reduced to 967 data records. Given that the values of 2, 2.5 and 3 are the common standard deviations values for detecting distortion, by considering these values and testing them on the data, the value 3 was used to determine the distortion data. Finally the number of data after removing duplicates and invalid data is reduced to 924 data records. The SPSSModeler.14 software is used for identifying outliers.

Then based on MIN-MAX method and using the formulas (1, 2, 3 and 4), indexes are normalized in the range of zero and one in order to prepare for modeling phase (next phase). In these formulas, MAXr, MAXf, MAXm, and MAXv represent the highest value, and MINr, MINf, MINm, and MINv represents the lowest value in the dataset, which are ultimately normalized to the final values of V', M', F', R'.

$$R' = \frac{(R - MIN_R)}{(MAX_R - MIN_R)} \qquad M' = \frac{(M - MIN_M)}{(MAX_M - MIN_M)}$$

$$F' = \frac{(F - MIN_F)}{(MAX_F - MIN_F)} \qquad V' = \frac{(V - MIN_V)}{(MAX_V - MIN_V)} \qquad (1, 2, 3 \& 4)$$

Phase 4. Modeling

In this phase, the proposed model is explained and developed. According to the purpose specified in the first phase, the CRISP- DM methodology as well as K-means algorithm that is one of the most famous and widely used clustering algorithm are used to develop the model. The K-Means algorithm is based on partition clustering that the number of clusters should be preset. Then, based on the number of initial clusters, data are placed in different clusters. One of the fundamental issues in this algorithm is to find the optimal number of clusters. In this study, the silhouette measure that uses the combination of two criteria of solidarity and density is used to determine the optimal number of clusters ranging from 3 to 10. In this

measure, the average distance of samples in a cluster is compared with the average distance of samples in other clusters. The results are indocated in table 2. According to the results, the best value for silhouette criteria in K-means algorithm is 3 clusters with the value of 0.540. Then, by increasing the number of clusters to a value of 10, standard numeric value of silhouette is declining. Whatever the output of this standard is closer to one, the quality of clusters resolution is better. Therefore, the optimal number of clusters will be 3. After that, the Mean value of behavioral variables in each cluster is evaluated. This helps to ensure that before proceeding to the next test, customers are grouped in significant clusters.

Table 2. Values of Silhouette Criteria in K-Mean Algorithm

	Cluster's Number	Value
1	**3**	**0.540**
2	4	0.491
3	5	0.492
4	6	0.461
5	7	0.468
6	8	0.476
7	9	0.445
8	10	0.448

- **Valuation and Determination of the RFMV Model Parameters Weight Using a Hierarchical Analysis:**

The weight of each parameter should be determined in order to rate and cluster customers. In this study, hierarchical analysis method, the most efficient decision support technique, is used for weighing the indexes of R, F, M and V. Then, respondents were asked to do paired comparison, and give the value of 1 to 9 to each index. The company's managing director and sales manager's ideas are used to determine the weight of parameters through the Expert choice software. According to formula (5), considering the total value of 1 and Inconsistency Index of 0.1, the total weight of 0.054, 0.075, 0.636, and 0.236 are obtained for the indexes of R, F, M and V respectively. Since the Inconsistency of variables is less than 0.1, the results are reliable.

$$WR + WF + WM + WV = 1 \quad (5)$$

Then, the indexes' value of each customer is calculated based on the formula (6):

$$CLVi = WR' * R' + WF' * F' + WM' * M' + W\,v' * V' \quad (6)$$

After determining the indexes' weight and value of all customers, customers' value is determined according to the Mean value of each cluster. Then, each index Mean is determined from the data extracted in the first phase (Table 3).

Table 3. Customers' Average Value

Cluster	Average Value	Total Average Value	Cluster Value Than the Average
1	0.33040	0.16718	Upper
2	0.07996		Lower
3	0.09155		Lower

Then, the total value of each cluster is divided by the number of members of that cluster, in order to determine the mean value of each cluster. The value obtained for each cluster is shown in Table 4.

Table 4. Clusters' Indexs' Mean

Cluster	$\overline{R}$'' Average	$\overline{F}$'' Average	$\overline{M}$'' Average	$\overline{V}$'' Average	Average Value
Cluster 1	0.0043	0.0384	0.1763	0.1113	0.3304
Cluster 2	0.0332	0.0027	0.0282	0.0154	0.0796
Cluster 3	0.0084	0.0110	0.0442	0.0276	0.0915
Total	0.0154	0.0174	0.0829	0.0606	0.1672

With regard to the main objective of the study, and given the value of customers based on behavioral variables, customers are categorized into three groups: high-value, middle-value and low-value. Due to the fact that each index can be dual-mode (above average and below average) the factors related to the variety of products are examined. The symbols ↑ and ↓ are used to compare the average of each cluster with total means. If the index is higher than the total average, the symbol ↑ is used and ↓ is used otherwise. The lower R (Recency) is better which means that customer purchased the products in closer intervals, while the value upper than mean is better for other variables. Given the customer value based on the behavioral variables, three categories of customers are considered in this study including, high-value customers that has better situation in terms of the average value of the indexes in the data, and indexes' mean in the cluster. This group has lower recency, and in terms of the other variables is better than other clusters. Moreover, the variety of products in the cluster is significantly higher and is distinct from other clusters.

Middle-value customers whose customers have an average level in terms of indexes, and have relatively low distance with the mean values obtained for indexes. In terms of the recency and variety of the products these customers are in the middle. In this study, nearly 51% of our customers are in this level. With regard to the number of members in the cluster, the movement of customers to other clusters is examined by dividing the cluster into sub-clusters. Firstly, for further investigation of the third cluster, this cluster is hierarchically divided based on K-mean algorithm.

According to the Means of R, F, M, and V variables, the Mean-values will be considered as a basis for of sub-clusters comparison. Based on the results, the highest silhouette is 0.454, which is related to the clustering with 5 clusters. Therefore, the data of third cluster is divided into 5 clusters based on K-Mean algorithm (Table. 5).

Table 5. Sub-Clusters Information

Variables (Higher than Mean-Value)	Total Percent	Number of Items	Sub-Clusters
FM	2.6%	12	1-3
RFMV	21.5%	100	2-3
MV	12.7%	59	3-3
R	37.1%	173	4-3
-	26.2%	122	5-3

The final group includes low-value customers. This group doesn't have a desirable condition based on the

indexes' values relative to the total value. The customers of this group have higher Recency, and lower Frequency and Monetary compared to the others. Moreover, the variety of purchased products is lower than, others.

Phase 5. Evaluation

After clustering and analysing the results, the differentiation of clusters created by K-mean algorithm and their resolution are evaluated through ANOVA test. As indicated in Table 6, significance levels (sig) of all variables (R (F=1.580, Sig= 0.000), F (F=790.748, Sig= 0.000), M (F=342.468, Sig= 0.000), and V (F=649.677, Sig= 0.000)) are lower than 0.05. Therefore, the homogeneity of populations' mean is rejected showing that clusters have different mean.

Table 6. ANOVA Test Results

Variable	Source of Change	SOS	df	Mean Square	F	Sig
R	Inter group	45.21	2	22.608	1.580	0.00
	Intra group	13.17	921	0.014		
	Total	58.39	923	-		
F	Inter group	24.23	2	13.117	790.75	0.00
	Intra group	15.27	921	0.017		
	Total	16.33	923	-		
M	Inter group	6.96	2	3.482	342.47	0.00
	Intra group	9.36	921	0.010		
	Total	16.33	923	-		
V	Inter group	20.82	2	10.410	649.67	0.00
	Intra group	14.75	921	0.016		
	Total	35.57	923	-		

After determining the incompatibility between variables, Post-Hoc Duncan test is used to ensure that clusters are distinct. In this test, if the variables do not have a significant distinction in the cluster, the similar values are placed in a sub-group. In other words, if we consider three clusters, the output of Duncan test should have three columns for each variable. The results of each variable are indicated in table 7. According to the results, the average values in all three clusters are quite distinct.

Table 7. Distinguishing Variable between Clusters

Cluster	variable	Number of Items	Clusters' Mean		
			1	2	3
1	R	181	0.803		
	F				0.512
	M				0.277
	V				0.471
3	R	466		0.160	
	F			0.147	
	M			0.069	
	V			0.117	
2	R	277			0.616
	F		0.038		
	M		0.044		
	V		0.065		

Phase 6. Deployment

After reviewing the results and valuation, the model is assessed. If the results are consistent with the primary targets of business, and seem to satisfy business needs, they will be used in a real environment. A new phase will be defined otherwisw, and the process will be repeated.

- **RFM and RFMV Comparison**

In this part, the results of the RFM and RFMV are compared with each other. Based on comparative analysis test, weights of R, F and M are equal to 0.088, 0.139 and 0.733 respectively in RFM model, while they are equal to 0.054, 0.075, 0.636 and 0.236 for R, F, M and V respectively in RFMV model. The average value obtained for the clusters is indicated in Table 8.

Table 8. Clusters' Average Value OF RFM & RFMV

Variable	Model	Cluster 1	Cluster 2	Cluster 3
Average value of R (Recency)	RFM	0.06	0.08	0.17
	RFMV	0.004	0.032	0.009
Average value of F (Frequency)	RFM	0.03	0.53	0.13
	RFMV	0.038	0.003	0.011
Average value of M (Montery)	RFM	0.04	0.26	0.07
	RFMV	0.176	0.028	0.044
Average value of V (Variaty)	RFM	-	-	-
	RFMV	0.111	0.015	0.028
Cluster's Average Value	RFM	0.09	0.28	0.09
	RFMV	0.330	0.080	0.092

Based on the results, in valuation using the RFM model the value of clusters is not distinct. Moreover, the values of cluster 1 and 2 are equal, therefore they are not distinct. However, the clustering based on a RFMV model by adding a variety of products (V), three clusters have different values. Thus, valuation would be better by using variable V. On the other hand, the role of the variable V in clusters' detachment and valuation is well confirmed. The results indicate that compared with clustering with RFM model, using the variety of products along with other behavioral variables provides more accurate clustering for companies.

4. Discussion

With the aim of finding groups with more product diversity along with other behavioral variables, the clusters are discussed in this section:

Cluster 1 (↑↑↑↑) high-value customers are the most valuable clusters among our customers and have the best condition in terms of the variety of the products. Moreover, regarding other factors are at a good level. This cluster has the potential to buy more and apply incentive programs, and is an appropriate group to impose the strategies of up-selling and cross-selling. The organization can make loyal customers by identifying and targeting these customers. Moreover, company can increase the probability of sales success and profitability by offering more products to this group.

Cluster 2 (↓↓↓↓) low-value customers are not in a desirable situation in terms of the obtained indexes. The notable point is that more than half of the firm's clients are in this cluster. This group has the least indexes' value. Low variety of the products of this cluster indicates that the customers of this group have low interest in diversified purchases. Therefore, offering additional products to these customers has not significant impact on the company's sales. Moreover, low recency of the cluster indicates its low frequency. This means that customers of this group for did

not interact with the company for a long time. The remoteness of the firm and the high recently can be one of the reasons for turning customer away from the firm.

Cluster 3 (↓↓↓↑) middle-value customers have an average value in terms of indexes. The recency of this group is less than average, which means the continuation of the customer relationship with the firm. Thus, with increasing volume and variety of products of this group, the monetary will increase. This group has a potentiality to become high-value customers through incentive programs and increasing the variety of products. The low average of this cluster in terms of monetary, frequency, and variety of products could be too risky. Low recency indicates that customers haven't interacted with the company for a long time. This will lead to the customer's defection. Since the cost of attracting new customers is high, company should provide appropriate strategies to keep its recency in a suitable level and to maintain customers, and then to increase the customers' value.

The sub-cluster of 3-2 (↑↑↑↑) has a higher average among other sub clusters of third cluster. Due to the high variety of products, this group has the potential to migrate to a higher level in the first main cluster. The group consists of 21.5% of total customers of third cluster.

Sub-cluster 3-5 (↓↓↓↓) is the most critical sub-cluster among others. The Average values of all the variables in this group are lower than total average values of the third cluster.

This group is on the verge of moving to the second main cluster, and then leaving the firm, and hence their disconnection. This group includes 26.2 % of total customers of third cluster that is a significant digit compared with other clusters.

Sub-clusters 3-1 (↓↑↑↓) and sub-cluster 3-3 (↓↓↑↑) are complementary. The organization could increase the variety of products and monetary factors of these groups to move into the sub-cluster 2-3, and then move to the first cluster of general classification by increasing the potential value of both groups using up-selling and cross-selling strategies. These two groups include about 15 percent of the third cluster's customers.

Sub-cluster 3-4 (↑↓↓↓) has kept its relationship with the firm, but their low volume of transactions is noticeable. The high recency of this group compared with the total mean is the only positive point of the group.

5. Conclusion

Social and economic changes between organizations and customers, have turned the transactional marketing to relational marketing. Although purchasing goods and services shape the nature of communication in organizations and businesses, few studies have considered the importance of this item. So, the nature of products is the missing link of researches in the field of customer behavior. In this study, the variety of products along with behavioral variables of RFM model was considered to identify the high-value customers. This would help organizations to find the best customers to implement the strategies of up-selling and cross-selling. Moreover, they can increase the customer buying domain through offering diverse products. The subsidiary purpose of this study is to compare the function of K-Mean algorithm on the data. In this study, the accurate identification of customers' groups as well as behavioral and demographic data was obtained using the RFM model besides adding the variable of variety of products to this model. Therefore, organizations can adopt the appropriate strategies in each category, and can increase their Probability of success. As the results indicated, the recency and frequency values in cluster 1 and 3 are pretty close together, but the significant differences in the values of variables F and R make a huge difference in CLV of these clusters. According to the results, variable V has the significant role in the distinction and valuation of clusters. On the other hand, the valuation using RFMV model is more distinctive than RFM model. Moreover, in clustering using the RFM model, despite the difference in variables' Mean, the average value of clusters 2 and 3 are equal, while are distinct in the RFMV model.

The results indicate that the variable V, variety of products, is effective in calculating customers' value. Moreover, from a practical perspective, it could have advantages such as accurate recognition of different groups, identifying subgroups to assess the movement of customers between groups, increased effectiveness of applied strategies, and increased profitability of organizations. As a whole, the results of modeling indicate that the variety of products along with other behavioral variables provide more accurate clustering than RFM model. It also could provide a more accurate understanding of customers' groups. This variable is important in providing campaigns and strategies that are directly related to the variety of products. Clustering based on RFMV model can specify the most distinctive and meaningful groups to apply up-selling and cross-selling strategies. In other words, the variety of products associated with each group can assist organizations to identify the target groups.

According to the results, the surveyed company that is B2B, can group customers in appropriate clusters by differentiating their characteristics.

Previous studies on the clustering customers using RFM model have only added one variable to the model. For more general information, to achieve more practical results, it is suggested that future studies add two or more variables to RFM model so to identify the variable with greater impact on understanding the customer. Furthermore, future studies can use methods such as Markov chain to study the movement of customers between groups. This will help organizations to analyze and evaluate the increase and decrease of customers in groups by combining and comparing the results. Moreover, it is suggested that future research expand the scope of study to various industries, and compare their results.

References

[1] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study". Procedia Computer Science, Vol. 3, 2011, pp. 57-63.

[2] P T. Kord Asiabi,, and R. Tavoli, "A Review of Different Data Mining Techniques in Customer Segmentation." Journal of Advances in Computer Research. Vol. 6, No.3, 2015, pp. 51-63.

[3] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. F. e Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences". Expert systems with applications, Vol. 39, No.12, 2012, pp.11250-11256.

[4] D. A. Kandeil, A. A. Saad, and S. M. Youssef, "A Two-Phase Clustering Analysis for B2B Customer Segmentation". In: Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on. IEEE, 2014. pp. 221-228.

[5] C. Liu, "Customer Segmentation and Evaluation Based on RFM, Cross-Selling and Customer Loyalty". In: 2011 International Conference on Management and Service Science. 2011.

[6] E. S. Margianti, R. Refianti, A. B. Mutiara, and K. Nuzulina, "Affinity Propagation and Rfm-Model for CRM's Data Analysis". Journal of Theoretical and Applied Information Technology, Vol. 84, No. 2, 2016, p. 272.

[7] H. Hwang, T. Jung, and E. Suh, "An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry". Expert systems with applications, Vol. 26, No. 2, 2004, pp. 181-188.

[8] C. F. Tsai, Y. H. Hu, and Y. H. Lu, "Customer segmentation issues and strategies for an automobile dealership with two clustering techniques". Expert Systems, Vol. 32, No.1, 2015, pp. 65-76.

[9] W. Wang, and S. Fan, "Application of Data Mining Technique in Customer Segmentation of Shipping Enterprises". In 2010 2nd International Workshop on Database Technology and Applications, on. IEEE. 2010, November. pp. 1-4.

[10] K. N. Lemon, and T. Mark, "Customer lifetime value as the basis of customer segmentation: Issues and challenges". Journal of Relationship Marketing, Vol. 5, No. 2-3, 2006, pp.55-69.

[11] B. Noori, "An Analysis of Mobile Banking User Behavior Using Customer Segmentation". International Journal of Global Business, Vol. 8, No.2, 2015, p. 55.

[12] Y. L. Chen, M. H. Kuo, S. Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data". Electronic Commerce Research and Applications, Vol. 8, No. 5, 2009, pp. 241-251.

[13] C. H. Cheng, and Y. S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory". Expert systems with applications, Vol. 36, No. 3, 2009, pp.4176-4184.

[14] J. T. Wei, S. Y. Lin, C. C. Weng, and H. H. Wu, "A case study of applying LRFM model in market segmentation of a children's dental clinic".Expert Systems with Applications, Vol. 39, No. 5, 2012, pp. 5529-5533.

[15] A. Asllani, and D. Halstead, "A Multi-Objective Optimization Approach Using the RFM Model in Direct Marketing". Academy of Marketing Studies Journal, Vol. 19, No. 3, 2015, p. 49.

[16] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty". Expert Systems with Applications, Vol. 37, No. 7, 2010, pp. 5259-5264.

[17] R. Venkatesan, and V. Kumar, "A customer lifetime value framework for customer selection and resource allocation strategy". Journal of marketing, Vol.68, No. 4, 2004, pp.106-125.

[18] M. Alvandi, S. Fazli, and F. S. Abdoli, "K-Mean clustering method for analysis customer lifetime value with LRFM relationship model in banking services". International Research Journal of Applied and Basic Sciences, Vol. 3, No. 11, 2012, pp. 2294-2302.

[19] W. J. Reinartz, and V. Kumar, "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing". Journal of marketing, Vol. 64, No. 4, 2000, pp. 17-35.

[20] R. A. Soeini, and E. Fathalizade, "Customer Segmentation based on Modified RFM Model in the Insurance Industry", Proceedings of 4th International Conference on Machine Learning and Computing IPCSIT. Singapore: IACSIT Press. 2012, Vol. 25, pp. 104-108.

[21] A. Ansari, and A. Riasi, "Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms". International Journal of Business and Management, Vol. 11, No. 7, 2016, p.59.

[22] H. C. Chang & H. P. Tsai, "Group RFM analysis as a novel framework to discover better customer consumption behavior". Expert Systems with Applications, Vol. 38, No. 12, 2011, pp. 14499-14513.

[23] D. T. Larose, Data mining methods & models. John Wiley & Sons, 2006.

[24] S. Moro, R. Laureano, & P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology". In Proceedings of European Simulation and Modelling Conference-ESM, 2011, (pp. 117-121). Eurosis.

[25] P. N. Tan, Introduction to data mining. Pearson Education India, 2006.

Saman Qadaki Moghaddam holds his master's degree in Information Technology Engineering from Islamic Azad University of Qazvin, Qazvin, Iran. He is interested in research in the field of information systems & technology, including innovation in ICT, and Electronic Commerce.

Neda Abdolvand is an assistant professor in Alzahra University and was a postdoctoral research fellow at Tarbiat Modares University. She holds her PhD and MS in information technology from Tarbiat Modares University and a post graduate certificate in information systems from Melbourne University. She is interested in research in the field of information systems & technology, including innovation in ICT, business intelligence, Electronic Commerce, and big data analytics. She has been a member of the Association for Information Systems since 2009.

Saeedeh Rajaee Harandi holds her master's degree in Information Technology Management from Alzahra University, Tehran, Iran. She is interested in research in the field of information systems & technology, including innovation in ICT, electronic commerce, and business intelligence.

Analysis of Business Customers' Value Network Using Data Mining Techniques

Forough Farazzmanesh (Isvand)
Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
foroghisvand@gmail.com
Monireh Hosseini*
Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
hosseini@kntu.ac.ir

Received: 15/Jun/2017 Revised: 12/Aug/2017 Accepted: 26/Sep/2017

Abstract

In today's competitive environment, customers are the most important asset to any company. Therefore companies should understand what the retention and value drivers are for each customer. An approach that can help consider customers' different value dimensions is the value network. This paper aims to introduce a new approach using data mining techniques for mapping and analyzing customers' value network. Besides, this approach is applied in a real case study. This research contributes to develop and implement a methodology to identify and define network entities of a value network in the context of B2B relationships. To conduct this work, we use a combination of methods and techniques designed to analyze customer data-sets (e.g. RFM and customer migration) and to analyze value network. As a result, this paper develops a new strategic network view of customers and discusses how a company can add value to its customers. The proposed approach provides an opportunity for marketing managers to gain a deep understanding of their business customers, the characteristics and structure of their customers' value network. This paper is the first contribution of its kind to focus exclusively on large data-set analytics to analyze value network. This new approach indicates that future research of value network can further gain the data mining tools. In this case study, we identify the value entities of the network and its value flows in the telecommunication organization using the available data in order to show that it can improve the value in the network by continuous monitoring.

Keywords: Business-to-business Marketing; Business Customers' Value Network; Market Segmentation; Data Mining; Telecommunication Industry; Value Network Analysis.

1. Introduction

As organizations move towards customer relationship management, the marketing function is the most impacted due to these changes. Under these conditions, data mining tools uncovering previously unknown patterns from large customer databases could help effective customer relationship management, because that can be done only based on understanding needs and preferences of the customers. However, discovered knowledge has to be managed in a systematic manner for true marketing [1].

Companies have different sources of value. Customers and their relationships with a company comprise an important part of organizational value [2]. So to survive in industry it looks building strong lasting relationships with customer who is possible through the value creation has become the most important marketing activity. Therefore, organizations must understand what creates customer value [3].

Value is the fundamental basis for all marketing activities [4]. From a company's viewpoint, all customers do not have the same value, and marketers evaluate customers to recognize key accounts for building a relationship with them [5]. In order to recognize key accounts many researches apply market segmentation using data mining techniques [6,7,8,9,10,11,12,13]. A concept guiding today's business in developing proper relation is the value of a customer defined as the profit resulting from customer's contribution. The common model for measuring customers' value is the RFM model, which made up of three major factors: recency, frequency, and monetary [14]. The model analyzing customer behavior [14, 15, 16] has been widely applied in many practical areas in a long history and has been the most frequently adopted in segmentation technique [16]. Recency represents the time of the last transaction, while frequency denotes the number of transaction in a certain period of time and monetary means the amount of money spent in this specified time period [14, 15, 16].

Nonetheless, large amounts of the value over time lost because many customers change their spending behavior more than defect so an important form of customer segmentation which to do is customer migration. It focuses on smaller changes in customer spending and considers customer value at different points in time to help companies not only stem the downward course, but also influence upward migration earlier [17,18]. Coyles and Gokey [17] found "that improving the management of migration as a whole by focusing not only on defections, but also on smaller changes in customer

* Corresponding Author

spending can have as much as ten times more value than preventing defections alone."

Marketing can be seen as network relationship interactions [19, 20] and network analysis can be used to describe the value creation [21]. In network perspective the key to value creation lies in understanding how value is created in relationships and it is the network of relationships that provides understanding the competitive environment. Therefore, we must extend any analysis away from viewing value creation. Dynamic nature is one of the most important aspects of network, because one relationship affects positively or negatively in others, an action by one participant in the network can influence other network members or an action by one participant may require further actions by other participants to be effectiveness [22]. In this regard, in order to help marketers recent studies introduced "customers value network", a new form of value networks an extension of Allee's [23] contribution, that focuses on relationships between a company and its business customers.

Customers' Value network is a mostly qualitative approach to distinguish business customers' different value dimensions and gain a better understanding of their customers. On the other hand, many organizations have enormous amounts of customer data in large databases. Marketers have realized that the knowledge in these huge databases is a key to supporting marketing decisions so use data mining tools to discover hidden knowledge about customers from those. However, data mining techniques haven't been used within the value network field, so we introduced a new multi-method for mapping customers' value network to gain the benefit from data available.

In this research following introducing research Methodology, we implement the approach in the case of a telecommunication for understanding the potential benefits. Finally, implications for marketers and related research are presented.

2. Methodology

On the one hand, Value network is an approach via which the great amount of qualitative information gathered through interviews link in a map to provide a fast, systematic way to analyze the data [24]. On the other, data mining is one important technique of knowledge discovery in the database which has been applied in many fields, especially in CRM and marketing in recent years. However, it has not been used in the value network field, so we proposed a novel multi-method to model customers' value network. A new model of value network applies data mining to use both qualitative and quantitative data.

2.1 Data Mining

Data mining is a process of extracting the unknown and valuable patterns from the huge amounts of data. The aim of data mining is to find out interesting knowledge from sizeable data set [25]. The identification of patterns in a large data set is the first step and Data mining tools provide marketers with just the customer knowledge but to gain useful marketing insights and making critical marketing decisions the discovered knowledge has to be managed in a systematic manner [1]. The process of data mining breaks into five major phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment, that the sequence of the phases is not strict and moving back and forth between different phases is always required [26].

2.2 Customers' Value Network

Value is the fundamental basis for all marketing activities [4] which the traditional mechanism creating it is the value chain [22, 23, 27]. In recent years, because this Mechanism is inadequate to understand the complexities of value exchange [22, 23, 27], that shifts to value network.

Value network is defined as" any purposeful group creating social and economic good through complex dynamic exchanges of value." Value network analysis is a business methodology that can help companies to distinguish and model value exchanges between networked parties as well as analyse, evaluate, and improve value conversion [21, 23, 27, 28]. This definition can be used in both internal and external perspective. External value networks include relationships between the organization and its suppliers, its investors, its strategic business partners, and its customers [21, 28].

Value networks involve different roles and organizations with different needs; hence, it is necessary to make specific propositions that create value for all participants in the network [29].

Traditional network research has extensively investigated the organizations that compose the network, while the whole network as a form of governance has not been so frequently studied [30]. Understanding network dynamics would influence managers' decisions regarding entering into new alliances by providing information on constraints from their current ties [31]. Network governance is needed for goal-oriented networks if they are to be effective [32]. Stable networks reinforce relational ties among members and ensure equitable distribution of value [33]. A new governance model is needed to realize a system in which 'sustainability issues are integrated in a way that ensures value creation for the firm and beneficial results for all stakeholders in the long term' [34,45].

The integration of sustainability at network level and the achievement of common and individual goals within the network could be then enhanced by new governance mechanisms. Rethinking the purpose of the firm as part of a value network could enable innovations towards new sustainable business models [38].

Greater stakeholder engagement is among the big changes that firms need to undertake in the pursuit of a long-term aim of sustainability [3637]. Den Ouden [29] suggests that specific arrangements are required for all parties in order to have a sustained portion both at the beginning and in the longer term, so they contribute to the

flourishing of the whole system. The analysis of the value flows within the network shows how different choices affect the mutual satisfaction of stakeholders, and hence the sustainability of the network [39]. Mutual value creation in sustainable business models, therefore, requires systemic consideration of a wide set of stakeholders who have a stake and responsibility in the value creation system [38].

In this regard, the study introduced a new concept of value network that focuses on relationships between a company and its business customers – called the "business customers' value network (BCVN)". Furthermore, they developed a systematic approach for mapping of BCVN that includes: describing the focal company, recognizing business customers, identifying the value exchanges, finalizing the value exchange and concluding the final map. To analyze business customers' value network, the paper used a combination of qualitative research approaches, namely in-depth interviews and consensus expert opinion. This approach helps marketers and managers distinguish business customers' different value dimensions in order to understand customers [40].

2.3 Proposed Method

Customers are the most important asset in each business unit, so organizations should develop long-term relationships with customers to survive in the challenging environment of a global market. To build a relationship with customers, organizations must understand its customers. However, on the one hand, hundreds of thousands of public customers have diverse dimensions of values and behaviors, and One-dimensional analysis isn't useful so it's essential to examine customers from several aspects. On the other hand, it is difficult to integrate and combine various analyses, so there is a need for a technique that summarizes them. An effective way to do this is visualization. Visualization is a technique for graphical representation of large amounts of data that its purpose is improving data representation to obtain maximum results and recognition.

In this section, a new approach for mapping and analyzing customers' value network is developed. Identifying and defining network entities of a value network, and drawing on innovative mapping, this approach presents an in-depth, multi-faceted picture of customers for marketing managers to gain a deeper understanding of their customers, the characteristics and structure of their customers' value networks and helps them better align their marketing strategies with the needs of customers to retain them. To conduct this work, we use a combination of methods and techniques designed to analyze large customer data-sets (e.g. clustering and RFM) and to analyze value network as well as to analyze customer migration. The method breaks into four major phases that are as follows and are shown in figure 1.

2.3.1 Business Understanding and Defining Network's Purpose

This phase includes the first step the process of data mining and BCVN. First, we begin by defining and describing the focal company. It is necessary to identify fully the aim of the company formation, its goals, its products and/ or services, its customers and its interactions with them. Next, understanding the objectives and requirements from a business perspective, we can define the purposes of network and its mapping (Indeed main aim of customers' value network is deeper understanding from customers).

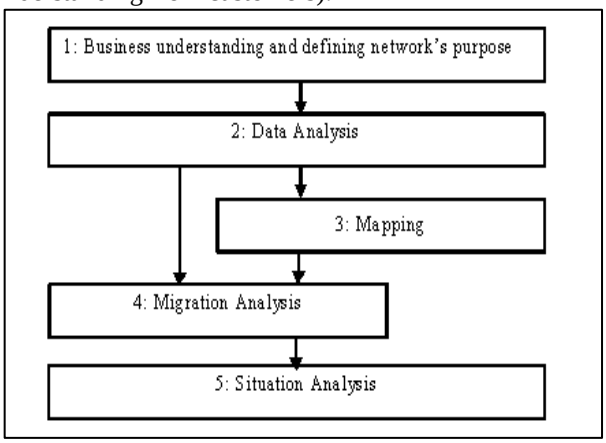


Fig. 1. Phases of proposed model

2.3.2 Data Analysis

In this phase, Data mining-related activities conduct in two steps, namely data preparation, modeling and evaluation. The data preparation phase covers all activities what to do to construct the final data set as data collection, data cleaning and so on. In modeling and evaluation phase, various modeling techniques must are applied and assessed to get objectives. To achieve favorable objectives can step back to the data preparation several times.

In preparation phase, First all customer data bases are identified then useful bases are selected and pre-processed. Second, as data variety is great, it's essential for achieving better analysis to classify data in categories according to similarity of them (as payment behavior, consumption behavior and so on). Third, the categories are prioritized using expert opinions.

In modeling phase, although various data mining techniques can be applied, the most important activity in our research is customer segmentation, because segmentation is a very common interest in marketing as well as we would primarily identify network entities. Therefore, customers with each category of features separately are clustered to determine segments of diverse aspects of customer behavior. Doing this, we understand how behavior of customers according to each category of features is. Results of the highest priority category are main segments, and others are Subsidiary.

2.3.3 Mapping

Mapping is a key phase in our approach. In this research, we only consider the relations between the focal company and its business customers (i.e. the star schema). In fact, for mapping network elements, there is no mandatory rule and it is more important that elements be selected in such a way that those get better understand. In order to be included mass information in a map, we should use from various elements and ways such as shape, shape division, color, page segmentation and so on. It is better that more significant segments are shown by more visible parts. Finally, we try to include important information in the map to draw a more intelligible picture from network. The basic mapping steps are as follows (it is essential to ٩* determine in any steps what element to use):

Step 1, identifying and describing Entities of networks using results of main segmentation

Step 2, identifying and describing subgroups of any Entity of networks using results of Subsidiary segmentation

Step 3, extracting creation patterns of Entity: we can discover the pattern using classification technique to can both analyze current patterns and monitor network changes over time

Step 4, grouping entities according to current value and future value (those can be guessed based on value segmentation.)

Step5, identifying value exchanges existing among nodes

Step 6, mapping

2.3.4 Migration Analysis using Data Mining Technique

Another form of customer analysis less noticed by many marketers is management of small changes in customer behavior while if not pay attention to that, they are losing a great deal of value. However, doing that is vital because continued tendency to downwards drift may causes significantly more value is lost over time than is lost through churn, and worse yet that customers tend not to return to their former group [17, 18]. It considers the customers group at different points in time to find out customers who change their buying patterns to give companies an early chance to correct any downward migration in their spending habits long before it leads them to defect [17]. Coyles and Gokey [17] found "companies not only can reduce downward migration by Systematic communication with target customers and tailoring tactics appropriately, but also helps them influence upward migration."

In this section, we apply data mining technique to analysis customer migration for gaining more insight of value network. The basic steps of migration segmentation are as follows:

Step1, identify value criteria of customer in order to do segmentation

Step2, Selecting a proper time interval for dividing the data set into two sections (to create data set1 and data set2).

Step3, implementing a clustering algorithm for data set1 for extract customer segments (the number of segments equal nodes of value network)

Step4, giving any segment a value degree

Step5, implementing classification algorithm based on results of clustering for identifying customer segments in data set2 (because patterns of segment formations within two data sets should be like to be comparable, we carry out this)

Step6, comparing segmentations (in all data, data set1 and data set2)

Step7, Clustering customers into three clusters:

- Positive migration: who have increased value degree
- Negative migration: who have decreased value degree
- Stable: who have held steady in their segment

2.3.5 Situation Analysis

Once the network mapped, we can now describe the situation of the customers' value network as it is. Describing network requires addressing several basic questions. The essential questions are:

- How are the overall patterns of creation entities?
- How are the overall patterns of migration?
- How are the overall patterns of exchanges and value creation?
- Does value management exist within the network?
- Is a group that received value from the organization is less than the value which that creates for the organization and it is necessary to be given it value-added?
- Is a group that be overheads (does not create any value for the organization)?
- What are the most appropriate channels to communicate with each group of customers?
- What are the opportunities of value creation?
- Which groups are appropriate for proposing new service?
- What are the overall patterns within the system? How assess you status of the network as a whole?

3. Case Study (Telecom)

In this section, in order to demonstrate the proposed approach application in the real world and to understand the potential benefits, it is applied to the Fixed-line Telecommunication Company. Step-by-step Implementation of the model is as follows.

Telecommunication operators are organizations that create value by linking customers who are interdependent. Value is about the connectivity to a network, which can be reached, and the conductivity of a network, what can be transacted [30]. This industry produce and storing a massive amount of data (i.e. call detail, network data) all over the world and because manual analysis of the extraordinary size databases is difficult, this industry is an early adopter of Data Mining technology. Telecommunication companies store a huge amount of data from and about its customers as Demographic data, Account Information, Service usage information, Bill information, Calling behavior information, payment behavior, etc. These data sources can be used for Comprehensive analysis of customers (analyzing all

aspects of customers). The application of data mining to customer relationship management can help telecom companies win new customers in the mass market, allow existing customers to be more profitable, retain valuable customers, offer their customers more personalized services, etc. [31].

Step-by-step implementation of proposed model in the fixed-line telecommunication is as follows. In the research, we applied the data set that includes data related to six two-month periods of Subscribers of Telecom Company of Dez (subset of KHZ Telecommunication). The company that has approximately 11000 fixed-line business subscribers provides services of voice call, ADSL, intelligent network (IN), easy internet (EI) and electronic payment system (1818). **"Discovering of creating-value opportunities for customer in order to generating more profits for the business" is defined as purpose of network.**

In second phase, in order to customer clustering, we used bill information that includes customer's monthly consumption and debt (i.e. data analysis is shown in figure 2). The two segmentation schemes considered are customer value segmentation and customer anti-value segmentation. Value segmentation focuses on identifying the contribution that a customer makes to overall organizational revenue based on current relationships with the organization. The other form of segmentation is segmentation according to customer debt. Because revenue is more important than debt from expert's view, main clustering is performed based on revenue information and debt information is used for the second level of clustering. Considering RFM model, we made of three major factors for value clustering: R, the time of the last revenue period, F, the number of revenue period, M, average of the amount of money spent in any period. As well as for anti-value clustering, three factors are made of: R, the time of the latest debt period, F, the number of debt period, M, average of the amount owed in any debt period. Using Two-steps algorithm, customers divide into five value segment and three anti-value segments that each reflects the different pattern of subscribers' behavior. Then according to segmentation results, other characteristics within each group as ARPU (Average revenue per user) and behavior service is gotten. Segmentation results are demonstrated in table1. Data mining is carried out using the default option of the spss Clementine 12 on a machine with RAM 8.00 GB and x64-based processor.

In the third phase, i.e. mapping, focal company is indicated by Oval and circles denote value segmentations discovered in the previous phase that the more customers in any group, the bigger its circle. Circle's labels indicate size and ARPU of any customer group. Any circle based on anti-value segmentation divided into sectors that show subgroups of that group, the arc length of each sector is proportional to the subgroup size. Color of any sector indicates its subgroup, and the darker be color, more degrees of anti-value (the same sector colour in different groups is alike). Value exchanges are indicated by arrows that can be bi-direction or one-direction (that bi-direction show Reciprocity). Furthermore, labels on arrows provide the description of value exchanges such as service name, Percentage of customers who use the services within the group and Percentage of the amount that the group has given for the service during the year. Finally, color of a circle header determines that group degree of value. Value segmentations based on current consumption behavior and future that guessed break up into four parts, namely platinum-colored, golden, silver-colored, bronze-colored. Figure 3 presents the mapped customers' value network.

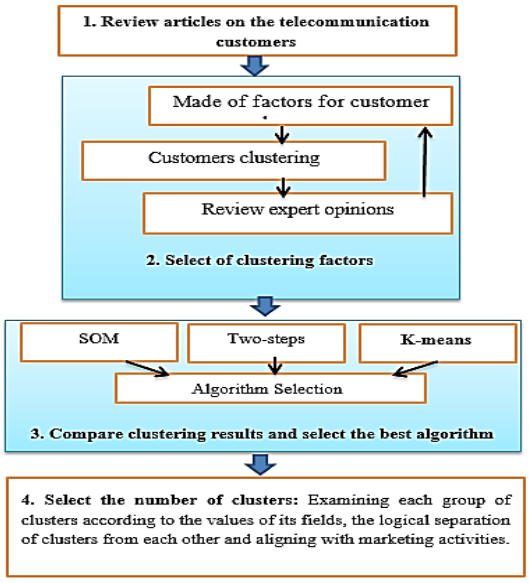


Fig. 2. Phases of data analysis

In fourth phase, i.e. migration analysis is performed using value segmentation. With the aim, data set is divided into two six-month part. So using two steps algorithm first-data set segment into five parts then implementing the classification algorithm second-data set segments are extracted. Next, segments are ranked from one to five. Finally, with calculating below formula, we identify migration segments: if the migration degree is more than 1, it's positive migration. If Migration degree is less than -1, it's negative migration. Else it's stable.

V: value degree of primary value segment, V_1: value degree of first-data set value segment, V_2: value degree of second-data set value segment, Migration degree: $(V_2 - V) + (V_2 - V_1)$.

Figure 4 presents statistics related to migration. Additional insights come from comparing the ratios of migration in any group. Subscribers become more stable by moving from left to right in figure 4 (Therefore, the more the debt is, the less stable it becomes). A substantial percentage of negative migration occurred in high and medium debt clusters; thus, it can be concluded that the debt of subscribers reduces the value of the network in the long term and makes the valuable clusters become smaller.

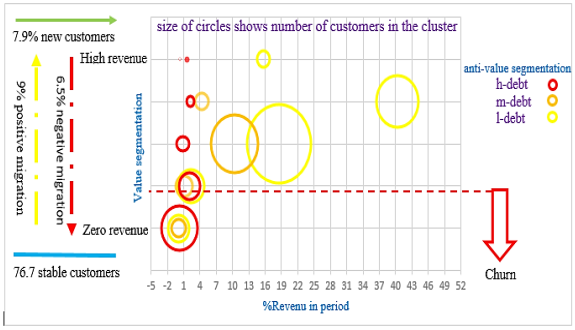


Fig. 3. Migration status

In fifth phase, i.e. situation analysis, we can describe the Situation of the customers' value network: according to the map the number of the most valuable subscribers is much lower than other categories. From high to low value groups, anti-value behavior is increasing. Network status is relatively bad because: The most valuable subscribers are very low, ARPU of middle groups its population is more than other groups is low, Many consumers have almost lost (this segments is shown as "churn customers" in figure 4 and in table 1), A substantial percentage of negative migration occurred that it is not a good sign. In the network value management do not exist because customers do not receive as much value from network that they get value to network. Unfortunately, Subscribers spite of disparate value receives equal services and facilities that the situation should not be continued. Also a group is that not only does not create any value but have cost for organization. As most service users are in the Middle group, they are appropriate for proposing new services and the opportunities of cross selling exist for higher group.

As we used various criteria for customer analysis, now any customer is part of a micro-segment. This allows more precision targeting, with knowledge of what the retention and value drivers are for each customer. Appropriate use of a multi-layer segmentation results in higher retention and growth. In addition, it allows enhanced business planning, where specific growth and retention targets can be assigned to every segment of each layer (namely value segmentation, anti-value segmentation, migration segmentation). Public strategies of any layer are in table 2.

Our approach uses both qualitative (namely in-depth interviews and consensus expert opinion) and quantitative data .Therefore, the accuracy of the results is taken from two dimensions .The algorithms and the method for examining the results of the quantitative are shown in Fig 2 and Qualitative results were also approved at meetings held with the organization's experts.

Table 1. Public strategies

Measures	**Level**		**Public strategy**
Main layer (value segmentation)	Platinum-colored Golden Silver-colored Bronze-colored	(high revenue) h-medium revenue medium and low revenue zero revenue	Retain retain and extend extend reduce cost
Second layer (anti-value segmentation):	Low-debt Medium-debt High-debt		encourage Should be encouraged provided that they pay its bill timely. Instalment payment
Migration layer:	Positive-migration Negative-migration Stable		to Strengthen and encourage upselling plan to prevent negative migration upselling plan

Table 2. Results of clustering

Low-debt				medium-debt				h-debt				
Count: 184,1.729 ARPU(rank): 6997636.125 (1)				Count: 4, 0.038% ARPU(rank): 6165947.5 (2)			Unstable in migration	Count: 18, 0.169% ARPU(rank): 5562774.889(3)			Unstable in migration	high-revenue
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
5.989	5.707	1211248		6.000	6.000	1027657		5.111	3.944	1324021		
4	1	750343		6	6	810947		2	1	737422		
6	6	6428363		6	6	1213734		6	6	3958889		
6	6	975183		6	6	1042975		6	3.500	909176		
R	F	M(debt)		R	F	M(debt)		R	F	M(debt)		
0.016	0.016	1722		3.250	1.250	229610		4.444	2.889	2748812		
0	0	0		2	1	79985		2	1	206766		
1	1	129452		4	2	498984		6	6	7853816		
0	0	0		3.500	1	169735		4	3	1725077		
Count: 1493,14.025 ARPU(rank):2255771.215(4)				Count: 187, 1.757% ARPU(rank): 1871379.316(6)				Count: 82, 0.77% ARPU(rank): 1925229.976(5)			Unstable in migration	medium- high -revenue
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
5.991	5.886	383636		5.957	5.701	330413		5.622	4.866	391538		
3	2	213245		4	2	213226		3	2	220854		
6	6	777914		6	6	759105		6	6	773020		
6	6	344216		6	6	283620		6	5	358016		
R	F	M(debt)		R	F	M(debt)		R	F	M(debt)		
0.042	0.042	3800		3.556	1.283	230869		4.573	2.683	836465		
0	0	0		1	1	52844		1	1	79543		
1	1	240595		6	3	584922		6	6	2436306		
0	0	0		3	1	196143		5	3	737211		

Low-debt				medium-debt				h-debt				
Count: 3382,31.771% ARPU(rank): 460002.847 (7)				Count: 1928,18.112% ARPU(rank): 446093.771 (8)				Count: 176, 1.653% ARPU(rank): 335949.028 (9)				
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
5.978	5.834	77806		5.959	5.780	76404		5.813	5.182	64457		
5	4	2982		5	4	11378		5	4	11378		
6	6	213118		6	6	213954		6	6	213641		
6	6	61086		6	6	67687		6	5	42983		
R	F	M(debt)		R	F	M(debt)		R	F	M(debt)		
0.055	0.055	5108		4.358	1.670	95571		5.352	4.017	300172		
0	0	0		1	1	50050		1	1	64444		
1	1	296306		6	3	565464		6	6	2777520		
0	0	0		5	2	77776		6	4	107955		
Count: 636, 5.975% ARPU(rank):133669.786(12)				Count: 292, 2.743% ARPU(rank): 178789.959(11)				Count: 443, 4.162% ARPU(rank): 204855.707(10)				
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
5.035	2.003	51671		4.589	2.349	50479		4.242	1.946	66775		
2	1	948		2	1	9525		2	1	10999		
6	4	463360		6	4	335884		6	4	589249		
6	2	16757		5	2	22579		4	2	21878		
R	F	M(debt)		R	F	M(debt)		R	F	M(debt)		
0.035	0.035	4673		4.682	2.024	108442		5.381	4.792	458573		
0	0	0		2	1	50695		1	1	62260		
1	1	271904		6	3	611278		6	6	7018612		
0	0	0		6	2	71088		6	5	267194		
Count: 446, 4.19% ARPU(rank): 64912.211(15)				Count: 213, 2% ARPU(rank): 71247.624 (14)				Count: 1161, 10.9% ARPU(rank): 71454.511 (13)				
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
0.029	0.029	1815		0.202	0.202	4752		0.028	0.028	2662		
0	0	0		0	0	0		0	0	0		
1	1	81046		1	1	213831		1	1	469961		
0	0	0		0	0	0		0	0	0		
R	F	M(revenue)		R	F	M(revenue)		R	F	M(revenue)		
0.087	0.081	1746		5.488	2.023	127302		5.993	5.885	391886		
0	0	0		1	1	60028		4	1	69582		
2	1	137582		6	3	796239		6	6	3954290		
0	0	0		6	2	110940		6	6	290714		
churn				churn				churn				
medium--revenue												
Low-medium-revenue												
Low -revenue												

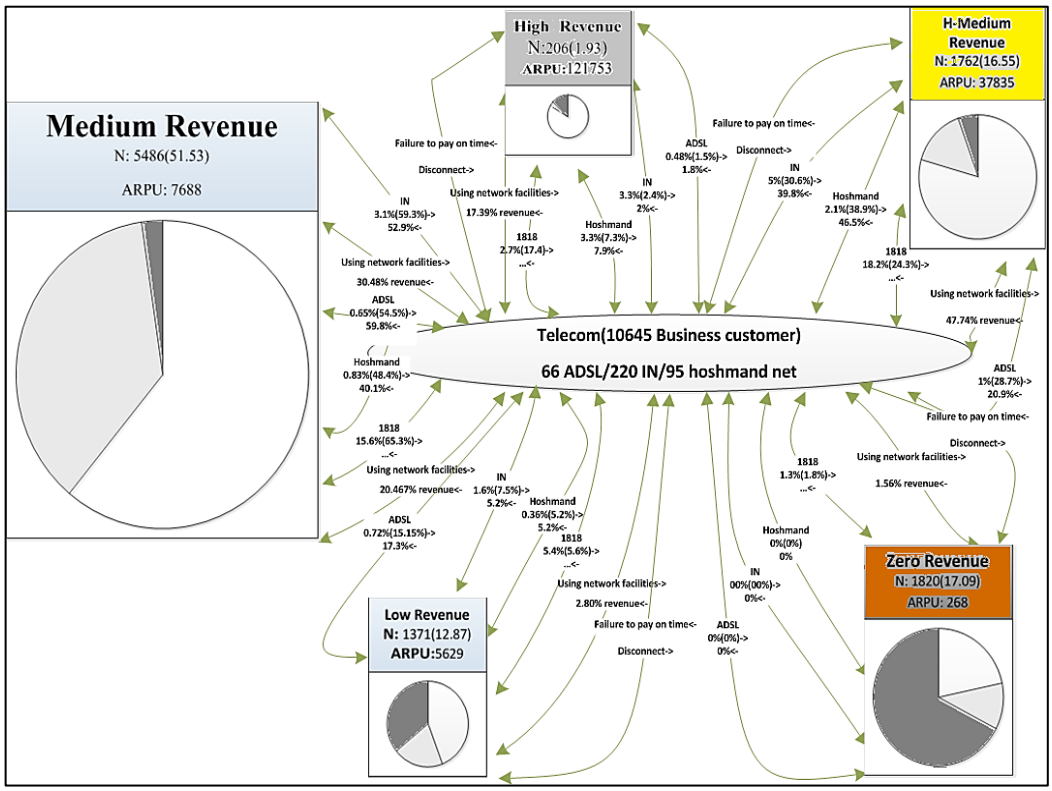


Fig. 3. customers' value network of Dez Telecom

4. Managerial Implications and Applications

Customers are the most important asset in each business unit but because of the intense competition and increased choices available for customers, organizations should manage customers in a long-term relationship. That requires that they interact with their customers based on actual customer preferences. As businesses owing to information technology increasingly have the capability to accumulate huge amounts of customer data in large databases, by using this databases data mining tools can help uncover the hidden knowledge and understand customer better. However, customers have diverse dimensions of values and behaviors and marketer for true understanding need to use multi-dimensional analysis (which each customer can be part of a micro-segment) to gain greater knowledge about customers. Since it's difficult to integrate and combine various analyses, there is a need for a technique that summarizes them. An effective way to do this is visualization. Data visualization allows marketers to view complex patterns in their customer data as visual objects complete in three dimensions and colors. Hence we proposed the approach that presents an in-depth, multi-faceted picture of customers for marketing managers to gain a multi-dimensional understanding of their customers so helps them better align their marketing strategies with the needs of customers to retain them. Additionally, because it requires no special technical knowledge for interpreting, it can be used as a managerial tool.

As regards the academic level, from the aspect of value network, we show future research of value network can further gain data mining tools. Our approach applies capabilities of data mining techniques in context of value network analysis. A summary of the differences between methodologies are in table2. In addition, from view of knowledge discovery, based on our results can understand that not only data mining tools help value network analysis but also value network map is very helpful for management of discovered knowledge using data mining. Indeed, value network can be considered as a visual framework for knowledge management that may be integrated into data mining.

In this case study, we distinguish parties (where any party is a group of business customers) and model value exchanges between networked parties using the available data in the telecommunication organization (focusing on the relationships between a company and its business customers). Then, diverse dimensions of values and behaviors parties are visualized in the value network to describe how the value is created in relationships. Analyzing value flows not only can prevent the decline of value in the network, but also it can identify ways to promote the value in the network.

In order to maintain the value of the network and increase it, managing the customer behavior changes in the network is a necessary task for a business and customer behavior changes in the network should be monitored continuously. Consequently, each category of migrations should be treated in such a way as to induce positive migrations behavior.

We also propose that it's better for organizations to investigate their value network in three levels: internal, customers and collaboration because a part of value lies in each of them. Figure5 shows different levels value network. And as Customers and their relationships with a company comprise an important part of organizational value [2], customers' value network is a very influential level.

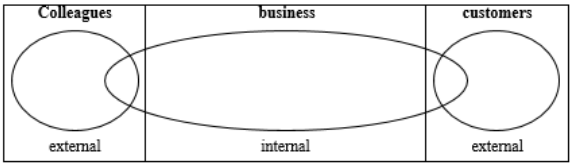


Fig. 4. Three level of value network

5. Conclusion and Limitation

Understanding the real value of customers is essential to retaining them. An approach that helps gain a deep understanding of customer is the value network by considering different value dimensions of customers. Despite enormous sources of data available in any organization, current approach of value network has not used them. This study, therefore, aims to justify the capabilities of data mining techniques in context of value network analysis. This study proposed a new approach to identify and define network entities in multi-layer in order to gain Deep insight about customer behaviors. Meanwhile, by understanding customers' value dimension, it allows precise targeting, where specific growth and retention targets can be assigned to every segment of each layer. Appropriate use of richly value network results in higher retention and growth. We applied this approach to the Telecommunications Company to show its applicability in the real world.

We in this research only analyze relations between the focal company and its customers and do not consider value relations existing among customers- call detail data can are used for this work. Moreover, we use a part of customer data, and a more detailed map can is created by using all of them. In future researches, we discuss how to use insights gained from this approach for decision making and policy management.

Acknowledgments

Authors are thankful to KHZ Telecommunication for providing data and acknowledge with grateful appreciation the kind assistance provided by Vice Chancellor for network affairs at the company, Mr. Naser Asad masjedi.

Table 3. Comparison of methods

Term	Value network (23) & BCVN [29]	Proposed approach
Method	This approach is a qualitative method that uses data gathered through interviews.	This approach is a multi-method analysis, because it applies data mining to use both qualitative and quantitative data. Hence its result surely will be more detailed. Quantitative means data that already have been in a company database.
Result	It is a one-level analysis that gives descriptive picture from nodes and its relation.	It is a multilevel analysis because both that gives descriptive picture with statistics and that uses diverse dimensions of values and behaviors. Therefore, the multi-faceted picture gets a deeper understanding of the network.
Mapping element	It uses specific elements for mapping, i.e. ovals, arrows, labels, solid and dashed lines.	It hasn't mandatory elements and can use various elements, such as shape, shape division, shape size, colour, and page segmentation etc.
Nodes	Roles (role-based network) or participants (participant-based network) in Allee's value network Business customers in BCVN	Business Customer groups
Node Attributes	Nodes can have attributes such as name and type.	As nodes identify using data, More detailed information is available. Nodes can have many identifiers such as name, size, ARPU and so on. Furthermore, color, division, position, etc. related to any node describes other characteristics of them.
Links	Link show value exchanges between two nodes and there can be multiple links between nodes.	Link show value exchanges between two nodes and there can be multiple links between nodes.
Direction	Every link indicates a single direction and bi-directional arrows are not permitted (in Allee's value network). The arrows should be one-direction, bi-directional arrows only are used to depict loyalty and sense of community (in BCVN).	Arrows can be bi-direction or one-direction that bi-direction show Reciprocity.
Link Attributes	Every link has unique attributes such as its deliverable and its nature (tangible or intangible).	Every link has unique attributes such as its deliverable and its related statistics that can be used as Evaluation indicator of that link.

References

[1] M.J. Shaw, M.E. Welge, S. Subramaniam and G. tan. "Knowledge management and data mining for marketing", Decision Support Systems, 31, 2001, pp. 127-137

[2] K. Hellman. "Strategy-driven B2B promotions", Journal of Business & Industrial Marketing, 20(1), 2005, pp. 4-11.

[3] Albert C. Socci. "Value-based Marketing for Bottom-line Success", Journal of Consumer Marketing , 22(1), 2005, pp. 50 – 51.

[4] M.B. Holbrook "The Nature of Customer Value", Sage Publications, Thousand Oaks, CA, 1994.

[5] L.Y.M. Sin, A.C.B. Tse and F.H.K. Yim. "CRM: conceptualization and scale development", European Journal of Marketing, 39(11/12), 2005, pp. 1264-90.

[6] H. Hwang, T. Jung, E. Suh. "An LTV Model and Customer Segmentation Based on Customer Value: a case study on the wireless telecommunication industry", Expert Systems with Applications, 26, 2004, pp. 181–188

[7] M.Y Kiang, M.Y Hu, D.M Fisher. "An Extended Self-organizing Map Network for Market Segmentation—a Telecommunication Example", Decision Support Systems, 42, 2006, pp. 36– 47.

[8] C. Mazzoni, L. Castaldi, F. Addeob."Consumer Behavior in the Italian Mobile Telecommunication Market, Telecommunications Policy", 31, 2007, pp. 632–647

[9] S.Y Sohn, Y. Kim. "Searching Customer Patterns of Mobile Service Using Clustering and Quantitative Association Rule", Expert Systems with Applications, 34, 2008.

[10] R. Hong, Z. Yan, W. Ye-rong. "Clustering Analysis of Telecommunication Customers", The Journal of China Universities of Posts and Telecommunications, 16(2), 2009, pp.114–116.

[11] I. Bose, X. Chen. "Exploring Business Opportunities from Mobile Services Data of Customers:An Inter-cluster Analysis Approach", Electronic Commerce Research and Applications ,9, 2010, pp. 197-208.

[12] P. Hanafizadeh, M. Mirzazadeh. "Visualizing Market Segmentation Using Self-organizing Maps and Fuzzy Delphi Method – ADSL Market of a Telecommunication Company", Expert Systems with Applications, 38, 2011, pp. 198–205.

[13] L.C. Cheng, L.M. Sun. "Exploring Consumer Adoption of New Services by Analyzing the Behavior of 3G Subscribers: An Empirical Case Study, Electronic Commerce Research and Applications", 11, 2012, pp. 89–100.

[14] S.H. Li, L.Y. Shue, S.F. Lee."Business Intelligence Approach to Supporting Strategy-making of ISP Service Management", Expert Systems with Applications, 35, 2008, pp. 739–754

[15] M.F. Bacila, A. Radulescu, I.L. Marar, "RFM based Segmentation: An Analysis of a Telecom Company's Customers", Marketing From Information to Decision, 5, 2012, pp. 52-62.

[16] J.T. Wei, S.Y Lin and H.H Wu. "A Review of The Application of RFM Model, African Journal of Business Management", 4(19), 2010, pp. 4199-4206, December Special Review

[17] S. Coyles, T.C Gokey. "Customer Retention Is Not Enough, Journal of Consumer Marketing", 22(2), 2005, pp. 101–105
[18] J. Bayer."Customer Segmentation in the Telecommunications Industry, Journal of Database Marketing & Customer Strategy Management", 17, 2010, pp. 247- 256.
[19] E. Gummesson. "Total relationship marketing: experimenting with a synthesis of research frontiers", Australasian Marketing Journal, 7(1), 1999, pp. 72-85.
[20] J. Zolkiewski and P. Turnbull. "Do relationship portfolios and networks provide the key to successful relationship management?", Journal of Business & Industrial Marketing, 17(7), 2002, pp. 575-97.
[21] V. Allee. "Value Network Analysis and Value Conversion of Tangible and Intangible Assets", Journal of Intellectual Capital, 9(1), 2008, pp.5-24.
[22] J. Peppard, A. Rylander. "From Value Chain to Value Network: Insights for Mobile Operators", European Management Journal, 24(2), 2006.
[23] V. Allee. "A value Network Approach for Measuring and Modeling Intangibles", Transparent Enterprise Conference, Madrid, 2002, available at: www.vernaallee.com.
[24] V. Allee, O. Schwabe. Digital Edition, "Value Networks and the True Nature of Collaboration", ISBN 978-615-43765-1, valuenet works and Verna Allee associates, 2011.
[25] Z. Ming. "Data Mining", Hefei: Press of University of Science and Technology of China, 2008.
[26] CRISP-DM. "The CRISP-DM Process Model for Data Mining", http://www.crisp-dm.org, 1999.
[27] V. Allee. "Reconfiguring the Value Network, Journal of Business Strategy", 21(4), 2000, pp. 1-6.
[28] V. Allee. "Value-creating Networks: Organizational Issues and Challenges", The Learning Organization, 16(6), 2009, pp. 427-442, Emerald Group Publishing.
[29] E. Den Ouden. "Innovation Design: Creating Value for People, Organizations and Society". Springer: London, 2012.
[30] KG Provan, A. Fish, J. Sydow. "Interorganizational networks at the network level: a review of the empirical literature on whole networks". Journal of Management 33(3), 2007, pp. 479–516.
[31] R. Gulati. " Alliances and networks", Strategic Management Journal 19(4), 1998, pp 293–317.
[32] KG. Provan, P. Kenis. "Modes of network governance: structure, management, and effectiveness", Journal of Public Administration Research and Theory 18(2), 2007, pp. 229–252.
[33] C. Dhanaraj, A. Parkhe. "Orchestrating innovation networks", Academy of Management Review 31(3), 2006, pp. 659–669.
[34] United Nations Environment Programme (UNEP)., "Integrated Governance: a New Model of Governance for Sustainability", Report by the Asset Management Working Group of the UNEP Finance Initiative, 2014.
[35] W.J.V. Vermeulen. "Self-governance for sustainable global supply chains: can it deliver the impacts needed?", Business Strategy and the Environment 24(2), 2015, pp. 73–85.
[36] R. Krantz, "A new vision of sustainable consumption. Journal of Industrial Ecology", 14(1), 2010, pp. 7–9.
[37] D. Bolton, T. Landells. "Reconceptualizing power relations as sustainable business practice". Business Strategy and the Environment 24(7), 2015, pp. 604–616.
[38] S. Evans, D. Vladimirova, M. Holgado, K. Van Fossen, M. Yang, E.A. Silva, and C.Y. Barlow. "Business Model Innovation for Sustainability: Towards a Unified Perspective for Creation of Sustainable Business Models". Bus. Strat. Env., 26:, 2017, pp. 597–608. doi: 10.1002/bse.1939.
[39] DR. Shaw. "Value creation in multi-level networks: a development of business model theory". In PACIS 2010 Proceedings, 2010, pp. 25–36.
[40] A. Albadvi and M. Hosseini. "Mapping B2B Value Exchange in Marketing Relationships: a Systematic Approach", Journal of Business & Industrial Marketing, 26(7), 2011, pp. 503-513.
[41] V. Allee, J. Taug, "Collaboration, Innovation and Value Creation at a Global telecom", The Learning Organization, 13(5), 2006, pp. 569-78.
[42] G. Asokan, S. Mohanavalli. "Fuzzy Clustering for Effective Customer Relationship Management in Telecom Industry", Springer-Verlag Berlin Heidelberg, CCSEIT 2011, CCIS 204, 2011, pp. 571–580 .

Monireh Hosseini holds a PhD from Tarbiat Modares University (TMU). She is currently an assistant professor at the Information Technology Department of Industrial Engineering Faculty at K. N. Toosi University of Technology. Her work deals with customer analytics and network models of customer value. She has teaching experience in Internet marketing, ecommerce strategies, electronic banking and Management Information Systems. She has published a number of research papers in international scientific journals and conference proceedings. She is the highly commended winner of the 2011 Emerald/EFMD Outstanding Doctoral Research Awards and received two scientific awards for the best paper in March 2008 and January 2011.

Forough Farazzmanesh (Isvand) was graduated in master degree of Information Technology Engineering at K. N. Toosi University of Technology. She received her B.S. degree in Information Technology Engineering from University of Isfahan. Her current affiliation is with the Iranian University department of computer engineering (An e-institute of higher education). Her research interest has focused mainly on data mining and its applications.

Concept Detection in Images Using SVD Features and Multi-Granularity Partitioning and Classification

Kamran Farajzadeh
Department of IT management, Islamic Azad University, Science and Research Branch, Tehran, Iran
k.farajzadeh@iau-tnb.ac.ir
Esmail Zarezadeh
Department of Electrical Engineering, Amir Kabir University, Tehran, Iran
zarezadeh@aut.ac.ir
Jafar Mansouri*
Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
jafar.mansouri@gmail.com

Received: 24/Oct/2016 Revised: 08/Aug/2017 Accepted: 20/Aug/2017

Abstract

New visual and static features, namely, right singular feature vector, left singular feature vector and singular value feature vector are proposed for the semantic concept detection in images. These features are derived by applying singular value decomposition (SVD) "*directly*" to the "*raw"* images. In SVD features edge, color and texture information is integrated simultaneously and is sorted based on their importance for the concept detection. Feature extraction is performed in a multi-granularity partitioning manner. In contrast to the existing systems, classification is carried out for each grid partition of each granularity separately. This separates the effect of classifications on partitions with and without the target concept on each other. Since SVD features have high dimensionality, classification is carried out with K-nearest neighbor (K-NN) algorithm that utilizes a new and "*stable*" distance function, namely, multiplicative distance. Experimental results on PASCAL VOC and TRECVID datasets show the effectiveness of the proposed SVD features and multi-granularity partitioning and classification method.

Keywords: High-Dimensional Data; Multi-Granularity Partitioning and Classification; Multiplicative Distance; Semantic Concept Detection; Static Visual Features; SVD.

1. Introduction

Semantic concept detection in images is the process of deriving meaningful terms that describe image contents. It is also referred to as image annotation [1] or indexing [2]. Semantic concept detection has been an active research topic in the recent years due to its potentially large impact on the image understanding, summarization, search, and filtering. It is essentially a classification task for determining the presence of the given semantic concepts in an image. The semantic concepts cover a wide range of topics such as those related to objects (e.g., car, airplane), indoor/outdoor scenes or locations (e.g., meeting, desert), and genre (e.g., weather, sports).

The success of a concept detection scheme strongly relies on the effectiveness of the low-level features in the content representation. Many systems have used global features mostly specified by MPEG-7 Visual part [3]. Global features include color or edge histograms, grid-based color moment and wavelet texture, etc. Other widely-used features are local features, like scale invariant feature transform (SIFT). Local features represent an image by the histogram of local patches based on a visual vocabulary of visual words [4]. An image is decomposed to a set of visual words derived after clustering or segmentation of the input image. However, local and global features have their own weaknesses. Global features do not contain local structure, and local features do not represent statistics about the overall distribution of texture or edge information. Moreover, most local features, like SIFT, are dependent on the size of the vocabulary of visual words; and the size of the vocabulary is also dependent on the type of images. For different types of data, the suitable size of the vocabulary changes.

This paper has the following contributions: New static visual features, namely, singular value feature vector, right singular feature vector, and left singular feature vector, are proposed. These features are derived by applying SVD "*directly*" to the "*raw*" images. In SVD features, different information like color, edge and texture is incorporated into an integrated framework and this information is sorted in accordance with their importance for detecting concepts. Moreover, feature extraction and classification is performed in a multi-granularity scheme, which is approximately similar to what is done by a human for detecting a concept. Classification is carried out for each grid partition of each granularity separately. Furthermore, feature vectors usually have high dimensionality even after dimension reduction. Recently, it has been shown that when dimensionality of data (feature vectors) is high, many distance functions (Minkowski distances, cosine similarity, etc.) become unstable [5][6][7][8]; i.e. distances

* Corresponding Author

of all data to a given query point become the same in the high-dimensional space. This phenomenon leads to the performance degradation of the classification algorithms that use these distance functions. To solve the instability problem, a new distance function, namely, multiplicative distance [8], is used that its stability in the high-dimensional space has been proved.

The rest of the paper is organized as follows. Section 2 gives an overview on the related works. Section 3 states characteristics of SVD features. This Section also introduces multiplicative distance. In Section 4, the proposed concept detection method is demonstrated in detail. The experimental results are reported in Section 5. Finally, some conclusions are drawn in Section 6.

2. Related Works

Generally speaking, two types of static visual features are often used: global and local. While global features are statistics about the overall distribution of color, edge or texture information, local features describe the local structures in an image. In the following, we mention some of these features in the previous works. Many papers like [9][10] have used global features, such as edge direction histogram, Gabor texture, color moment, color histogram, canny edge, etc., for describing images. Most of these features are defined by MPEG-7 Visual part. These features either are concatenated to form a single feature vector for the classification [9] or are used separately for the classification [10]. In both of these cases, the relationships between the features are not usually taken into account. In [11] local binary pattern (LBP) has been used for image feature description. In [12] an image descriptor based on the orientation of contrasts is proposed. The contrast and width of canny edge features are computed by a simple scheme. Inspired by the histogram of oriented gradients method, an image representation is proposed based on a histogram of contrasts.

Some further research works fall in the category of the local features. An image has local interest points or keypoints defined as salient patches that contain rich local information about the image. Keypoints are usually around the corners and edges of image objects, such as the edges of the map, people's faces, etc. The most popular keypoint-based representation is bag-of-visual-words (BoW). In BoW, a visual vocabulary is generated through grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. An image can be represented as a histogram of visual words. The performance of BoW features in semantic concept detection is subject to various representation choices [4].

Lazebnik et al. [13] have exploited the spatial location of keypoints and proposed a spatial pyramid matching (SPM) method, in which an image is first divided into multi-level equal-sized grids and each grid is described by a separate BoW using SIFT descriptor. Then, the BoWs from image grids at each level are concatenated to form the final representation. Fisher vector is another type of the image representation that has been used in [14]. The Fisher vector can be seen as an extension of the BoV. Both of them are based on the visual vocabulary built on the low-level features like SIFT descriptor. If a Gaussian mixture model is used to model the visual vocabulary, the gradient of the log likelihood can be computed with respect to the parameters of the model to represent the image. Sparse coding is a feature encoding method that has been widely-used in recent years [15]. The aim of sparse coding is to represent input vectors as a linear combination of a small number of basis vectors (dictionary).

In [16] a method called non-negativity and locality constrained Laplacian sparse coding is proposed. Firstly, non-negative matrix factorization is used in the Laplacian sparse coding, which is applied to constrain the negativity of both codebook and code coefficient. Secondly, K-nearest neighboring codewords for local features are used because locality is more important than sparseness. Finally, non-negativity and locality constrained operators are utilized to obtain a novel sparse coding for local features. Convolutional neural network (CNN) has been used in [17]. It consists of multiple convolutional layers of small neuron collections followed by fully connected layers. These layers form multi-stage feature extractors, which higher layers generate more abstract features from lower ones. The input to the CNN is raw image pixels such as an RGB vector, which is forwarded through all feature extractor layers to generate a feature vector that is a high-level abstraction of the input data.

Some papers have used combination of global and local features. Jiang et al. [4] have used local BoW feature and two types of global features, color moment and wavelet texture, which have been obtained from the whole image. The local and global features have been combined and classification has been performed for the whole image, neither for each partition nor for each granularity.

3. Preliminaries

3.1 Motivations for Using SVD Elements as the Low-Level Feature

SVD elements can be useful for expressing edges, textures and colors/luminance in images. For describing this matter, notice that an image, an $m \times n$ matrix A, can be interpreted as the ensemble of the basis images as follows:

$$A_z = \sum_{i=1}^{z} \sigma_i u_i v_i^T \quad (1)$$

Where $z(z \leq p, p = \min(m,n))$ is the number of u_i (left singular vector) and v_i (right singular vector) pairs used. Each $u_i v_i^T$ specifies a layer of the image geometry, whereas the singular value σ_i is the weight assigned to this layer and specifies the luminance of that image layer [18][19][20]. The first few singular vector pairs account for the major image structure, whereas the subsequent u_i and v_i pairs account for the finer details in the image.

Larger singular values indicate more energy in the image. The singular values also denote the activity level in the image. A high activity level represents roughness or strong textures and edges. Similarly, a low activity level corresponds to smoothness or weak textures and edges [18][20]. These show that SVD can be used for representing different regions of an image.

If the first few singular values have predominant magnitude, after projecting along the first few singular vector pairs the reminder scatters are small and ignorable. This point is demonstrated through an example shown in Figure 1, where the image size is 352×288 and thus p=288. It can be observed that the first 30 basis images (z=30) [i.e., i = 1 to 30 in (1)] capture the major image structures and luminance/colors, and the subsequent basis images signify the finer details in the image. Furthermore, it is observed that with some few singular values and their respective left and right singular vectors, an image can be reconstructed nearly similar to the original image.

(a) original (b) Z =10 (c) Z =20 (d) Z =30 (e) Z = 40 (f) Z = 50

Fig. 1. A_z as defined by (1) for different values of Z, (a) original image ($Z = 288$), (b) Z= 10, (c) Z = 20, (d) Z = 30, (e) Z =40, (f) Z =50.

In SVD, color/ luminance, texture, and edge information is sorted according to their significance. SVD integrates this information simultaneously and takes into account the relation between them. Hence, all the edge, color and texture information is encoded into a single representation. In addition, there is no redundant information in SVD since left and right singular vectors are orthonormal. Furthermore, SVD takes into account human visual perception [20].

Moreover, singular values are stable. The stability of singular values indicates that when there is a little disturbance in the image, singular values do not change considerably [18]. Therefore, singular value features can be effective to encounter noise, small clutters and small changes in the image. Additionally, singular values are useful when the image has transposition, rotation and translation; since a) a matrix A of an image and its transpose, A^T, have the same non-zero singular values [18]; b) if R is a unitary and rotating matrix, the singular values of RA (rotated matrix) are the same as those of A [19]; c) the original image A and its rows or columns interchanged image have the same singular values [19]. These abilities motivate us to use SVD elements as the low-level feature for the concept detection.

3.2 Multiplicative Distance

Under some conditions on the data distribution, distances between data and query points in the high-dimensional space are meaningless or unstable [5][6][7][8]. This means that distances of all data from a given query point become the same for a wide variety of data distributions and distance functions, when dimensionality increases toward infinity. Minkowski and fractional norm distances [8], cosine similarity for the i.i.d. (independent and identically distributed) data [21] are some examples. In such cases, the concept of proximity and similarity is not meaningful because of the poor discrimination between the nearest and furthest neighbors. This instability can greatly affect many applications like classification, and can result in the performance degradation. In [6] and [7] authors have stated that the sufficient and necessary condition for the stability of a distance function. In [8] a new distance function, namely, multiplicative distance, has been introduced that its stability has been proved. The multiplicative distance function can be used in the low-dimensional space. The definition of this distance function is as follows.

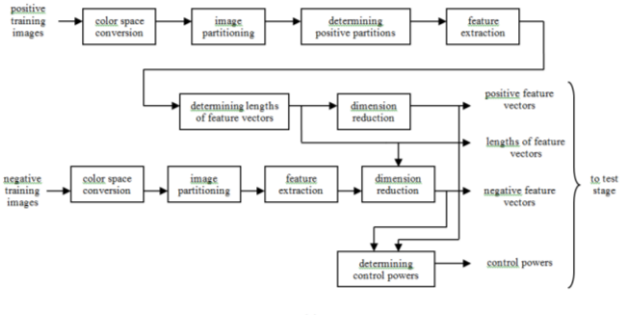
Definition: Let $X = (x_1, x_2, \dots, x_m)$ be an m-dimensional random vector with $x_k \sim F_k$ (F_k is the distribution of the random variable x_k), $k = 1, \dots, m$ and $Q = (q_1, q_2, \dots, q_m)$ be the query point with $q_k \sim \widetilde{F_k}$. Set $z_k = 1 + |x_k - q_k|$. The general form of the multiplicative distance of X from Q is defined as:

$$\mathrm{MD}(X,Q) = \left(\prod_{k=1}^{m} z_k^{c_k} \right) - 1 \qquad (2)$$

where c_k is named "control power", which controls the effect of each z_k on the distance. z_k^{ck} is defined as distance component. If $\forall k: c_k = c$, each dimension has equal effect on the distance. In the simple form of the multiplicative distance we have $\forall k: c_k = 1$.

4. Proposed System

Figure 2 shows the block diagram of the proposed concept detection for each case of multi-granularity partitioning, which includes training and test stages.



(a)

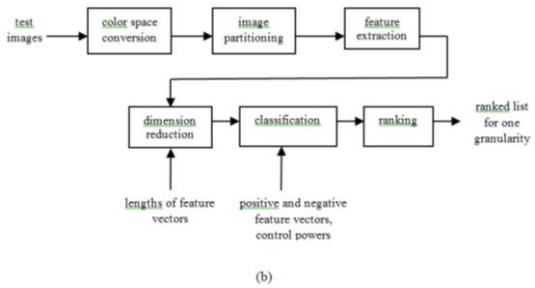


Fig. 2. Block diagram of the proposed concept detection for one case of multi-granularity partitioning, (a) training stage, (b) test stage.

4.1 Training Stage

First, positive images (i.e. images containing the target concept) are partitioned into different granularities. Usually, six cases of granularities, i.e. 1×1 grid (1 partition), horizontal 2×1 grid (2 partitions), vertical 1×2 grid (2 partitions), 2×2 grid (4 partitions), 3×3 grid (9 partitions) and 4×4 grid (16 partitions) are used for feature extraction for most of concepts. Since some concepts are depicted by the holistic representation of an entire image rather than a region (e.g., office), very small granularities (like 4×4 grid) are not used for these concepts.

This kind of partitioning is approximately similar to what is done by the human for detecting concepts. When a human looks for a concept in an image, in accordance with the concept, first, he/she looks at the whole image and if it is necessary, the view range is getting smaller in order to detect the concept. So, the human hierarchically narrows down the looking range until detecting the target [22][23]. However, the human perception system is very complicated and uses extensive rules. In the proposed method, the similar work is done but in a simple way. The positive images are divided into equal grid partitions at different granularities so that they are perceptible for the human for detecting concept. Additionally, very small partitions are not used since they have no perceptual meaning.

Different types of partitions based on multi-granularity partitioning are obtained. Then, partitions that contain the concept (positive partitions) are detected manually and are kept. As stated in Section 3, SVD elements can be used as the low-level features. Therefore, SVD is applied to each positive partition of the image. SVD features are extracted in three cases: from "*raw*" color images with RGB and HV color spaces and "*raw*" gray scale space. The right singular vectors of SVD are concatenated to form the right singular "feature" vector. The left singular vectors of SVD are also concatenated to form the left singular "feature" vector. Singular values are also put in the singular value "feature" vector. Therefore, in the case of experiments with the gray scale space, there are 3 feature vectors for each partition; and for the case of experiments with the HSV and RGB color space, there are 9 feature vectors (3 feature vectors for each component of the color space).

Notice that SVD features are obtained "*directly*" from "*raw images*" of partitions. Our SVD features are different from features in methods that use techniques like PCA or LSA. In those methods, low-level features (like SIFT features) are extracted. Then, SVD in techniques like PCA or LSA is applied to the "*extracted low-level features*" (not on the "*raw images*") to produce the final processed features.

SVD features are obtained from the positive partitions. However, feature vectors usually have very high dimensionality. High-dimensional data requires large storage space and more computation and to reduce the computational and storage cost, dimension reduction is necessary. With the dimension reduction, some data that do not help for detection are removed and performance can be improved. In addition, the dimension reduction is useful for the noise cleaning. An interesting property of SVD is that information is sorted based on its importance in the descending order. Small or zero singular values indicate that their respective right and left singular vector pairs (in Equation (1)) have less or no significance. Furthermore, if the first few singular values have a predominant magnitude, after projecting along the first few singular vector pairs, the remainder scatters can be ignored. Therefore, with removing less or insignificant singular values and vectors, the dimensionality is reduced. For this task in the training stage for each positive partition (of each case of granularities) that contains the target concept, its energy is calculated from the below formula:

$$E = \sum_{i=1}^{p} \sigma_i^2 \tag{3}$$

where $p = \min(m,n)$, m and n represent the size of the partition. Then, the first index of the singular values that satisfies the following equation is obtained:

$$E_{th} = th \times E \le \sum_{i=1}^{index} \sigma_i^2 \tag{4}$$

where *th* is a threshold. E_{th} is the energy of the reconstructed image (partition). *th* is selected so that after the dimension reduction, the reconstructed image has good perceptual quality and no considerable distortion. For simplicity, *th* is selected the same for all concepts (for all granularities and also all components of color images). For each case of partitioning these indices are calculated separately for all partitions containing the target concept, and the final index for each granularity, which is used in test stage, is the average of these indices. Thus, we have different indices for different granularities (e.g., 6 final indices for each kind of feature vector in the case of 6 granularities with gray scale images). For HSV and RGB cases, for each color component of the images this process is performed separately. Therefore, usually different indices related to *th* are yielded for three components of the color images. These final indices determine lengths of feature vectors for the concept. With removing the singular values after these indices and their respective elements in the left and right singular feature vectors (i.e. their respective left and right singular vectors in Equation (1)), the dimensionality of feature vectors are reduced and final positive feature vectors are yielded. The reduced-dimension feature vectors of the proposed system have different lengths for different concepts.

Based on Figure 2, for the negative images (images without the target concept), multi-granularity partitioning is performed. Then, SVD features are obtained from partitions. Using lengths of feature vectors (from obtained indices), the dimension reduction is carried out on these features and the final negative feature vectors are attained. Furthermore, since for the classification in the test phase the multiplicative distance is used, the parameter c (control power) for the multiplicative distance is selected such that the overflow does not happen. For this purpose, for each granularity and for each type of feature vector for each concept, the pairwise multiplicative distances of all feature vectors (positive and negative) for different values of c are calculated. The largest value of the non-positive integer powers of 10, i.e. 1, 0.1, 0.01, ..., for which overflow does not happen, is selected as the control power. The final positive and negative feature vectors, lengths of feature vectors and control powers are used in the test stage.

4.2 Test Stage

For each test image, multi-granularity partitioning is done. Then, SVD is applied to the partitions of raw images and SVD features are obtained from each partition. The dimension reduction is performed on the feature vectors of each partition using the lengths of feature vectors (final indices from the training stage). However, even after dimension reduction the feature vectors usually have high dimensionality. As stated in Section 3.2, conventional distance functions in the literature become unstable for the high-dimensional data. So, using these distances in a classifier can result in the performance degradation. Multiplicative distance has been introduced as a stable distance function [8] and we use this distance in our work. Notice that a classifier must be used that this distance function is applicable to it. For this reason, classification is carried out with the well-known K-NN algorithm with the stable multiplicative distance function.

If we consider features of all partitions of an image together for the classification, some partitions may not have the target concept. Therefore, wrong detection can occur due to considering features of the partitions with and without the target concept together. Thus, in the proposed method classification is carried out for each grid partition of different granularities individually. Furthermore, classification is performed for each of 3 or 9 kinds (based on gray scale or color images) of feature vectors of each partition separately.

It should be noted that our multi-granularity partitioning and classification system is different from the spatial pyramid matching approach, commonly used in BoW representation like in [13]. In [13] an image is partitioned into different granularities. For each granularity, BoW features are obtained. Finer granularities get higher weights. Then, features of all granularities are concatenated and classification is performed for the whole image not for each partition. In our work in contrast to [13], features of different granularities are not concatenated and detection is performed for each grid partition of each granularity separately. Furthermore, different granularities have the same importance in detection. This is because finer granularities do not necessarily represent concepts better. Moreover, for some concepts we do not use small granularities as stated before.

If each feature vector has n dimensions and number of the training feature vectors is m, and $X_i = (x_{i,1}x_{i,2}, \dots, x_{i,n}$ $i = 1, \dots, m$ are the training feature vectors, and $Y = (y_1, \dots, y_m)$ is a feature vector of the test sample, the classification is as follows:

$$label \quad Y = label \quad \arg \min_i \prod_{j=1}^{n} \left(|x_{i,j} - y_j| + 1 \right)^c$$

$$f(Y) = \min_i \prod_{j=1}^{n} \left(|x_{i,j} - y_j| + 1 \right)^c \qquad (5)$$

where f is the distance output of the classification. *label* of Y is positive if X_i that has the minimum distance is from the positive training feature vectors; otherwise *label* is negative. C is the control power of the multiplicative distance. For all 9 (for RGB and HSV images) or 3 (for gray scale images) feature vectors of a partition, the classification is performed separately. For a partition, if at least one of the classifications on its feature vectors gives positive answer for the target concept, that partition is annotated with the positive label. For an image, if at least one of its partitions is positive, that image is annotated with positive label for that granularity.

For each concept, following stages are performed for "each granularity":

a. For each positive image the partition(s), namely, the best partition (s), with the maximum number of positive labels of the classifications is (are) kept as the representative of that image and other partitions are eliminated.
b. The best partitions of positive images are divided into 3 or 9 groups in a descending order in accordance with the number of their positive labels. For example, for the case of gray scale images, there are 3 groups: the first group is related to the best partitions that have 3 positive labels; the second group contains the best partitions with 2 positive labels; and the last group is related to the best partitions with 1 positive label.
c. Among each group of best partitions, for each "kind" of feature vector, feature vectors (with positive or negative label) are ranked based on their labels and distances, *f*. First, positive feature vectors are ranked in an "ascending" order according to their distances. Next, negative feature vectors are ranked in a "descending" order based on their distances. The score of each best partition is the summation of ranks of all its feature vectors.
d. If an image has some best partitions (i.e. with the same number of positive labels), the partition with the best score is kept and other partitions of that image are eliminated.
e. The best partitions are ranked according to their group and score, to form the ranked list of positive

images for that granularity. Note that each best partition is related to one image.

Next, the ranked lists of all granularities are aggregated with simple fusion to form the final ranked list of positive images for that concept.

5. Experimental Results

In this Section, the performance of the proposed method is evaluated on two well-known PASCAL VOC and TRECVID datasets. The first dataset is the widely-used PASCAL VOC 2007 [24], which consists of 8 concepts and includes 9963 images divided into a predefined training and test set of 5011 and 4952 images, respectively. Figure 3 shows the example images for 8 concepts in PASCAL VOC 2007.

Fig. 3. Exemplary images for the evaluated concepts in the experiments in PASCAL VOC 2007 dataset.

The images are resized to 352×288. For each semantic concept, average precision (AP) is used for the performance evaluation. AP is the average of precisions computed at the point of each of the relevant (by the ground truth annotation) images for considering the order in the ranked list of images [25]. To evaluate the overall performance, we use mean average precision (MAP) which is the mean value of the APs over all concepts.

The second dataset is the keyframes of TRECVID 2007 (TV07) [26]. The national institute of standards and technology (NIST) has established "semantic indexing" as a task in TREC video retrieval evaluation (TRECVID) [26], which aims to provide a benchmark for evaluating video concept detection technologies. The shot and sub-shot detection and extraction of keyframes have been performed by TRECVID. The training and test datasets consist of 21532 and 22084 keyframes, respectively. The keyframes are in CIF (352×288) format. There are 20 semantic concepts evaluated in TV07. Figure 4 shows the example keyframes for all 20 concepts evaluated by TRECVID. The reason for selecting TV07 is that all concepts are detectable by visual features.

Fig. 4. Exemplary keyframes for the evaluated concepts in the experiments in TRECVID 2007 dataset.

For each concept, inferred AP average precision (infAP) is used for the performance evaluation. InfAP is designed for partially labeled datasets (like TV07 test set). The infAP is an approximation of the AP and can save significant judging effort during the annotation of ground truth for large test dataset [25]. For each concept, infAP is computed based on the returned rank list and the ground truth provided by TRECVID. Notice that for the semantic concept detection in TRECVID, since the final annotation has been carried out for shots, if one shot has some positive sub-shots, just the sub-shot whose keyframe has the better rank is kept and other sub-shots are removed. Note that each keyframe is related to one sub-shot. Following the TRECVID evaluation, the infAP is computed over the top 2000 ranked shots according to the outputs of the proposed system. The mean infAP (MinfAP), which is the mean value of the infAPs over all concepts, is used for evaluating the overall performance.

First, we compare the proposed multi-granularity partitioning and classification method, which is referred to as MGPC, with the spatial pyramid matching (SPM) method [13] that is the most well-known partitioning and classification scheme and similar to our method. In SPM an image is first divided into multi-granularity equal-sized

partitions and each partition is described by a separate BoW using SIFT descriptor. Then, the BoWs from image partitions at different granularities are concatenated with weights proportional to the level of the granularities to form the final representation for that image. Notice that bigger granularities (higher levels) have higher weights. For having a fair comparison, for both of the MGPC and SPM, the SIFT descriptor is selected as the low-level feature and two classifiers, the K-NN algorithm with the multiplicative distance and SVM with RBF (radial basis function) kernel are used for both methods. Notice that validation experiments have been performed for selecting K in K-NN (for K=1,3,5,7,9) and parameters C and gamma in SVM (C= $2^{-5}, \ldots, 2^{4}$ and $\gamma = 10^{-7}, \ldots, 10^{2}$). For brevity of the paper, just the final results are reported. Based on the experiments, for K-NN, K=5 and for SVM, C=4 and $\gamma = 0.1$ are obtained. The selection of control power in the multiplicative distance has been illustrated before in the training stage.

It is important to note that partitioning is continued until partitions are perceptible for human detection of a specific concept. For PASCAL VOC dataset, all 6 granularities are used for all concepts. But for TV07 dataset, some cases of granularities are not used for some concepts. For "people_marching" and "meeting", 4×4 and 3×3 grid cases and for "office", 4×4, 3×3 and 2×2 grid cases are not used. This is because it seems that partitions of these cases cannot represent these concepts individually. For example, one partition in 3×3 grid case usually cannot represent the concept of "office" in an image. Table 1 presents the definition of fusions, i.e. number of granularities used. For example, for MGPC method, fusion5 means that 5 granularities are used for partitioning. Notice that for the SPM method, definition of fusions in Table 1 means the "levels of spatial pyramids" used in [13]. Figures 5 and 6 show the performance of the proposed method and SPM method for PASCAL VOC and TRECVID datasets for different fusions.

From Figures 5 and 6 it is observed that the proposed method has the superior performance over the SPM method. The reason is that in SPM, features of partitions of different granularities are combined to form the final feature vector. Then, just one classification is carried out on the final feature vector. Thus, features of the partitions "without" the target concept can affect features of the partitions "with" the target concept and this can lead to the wrong result in the classification. However, in the proposed method the above problem does not occur. The reason is that the classification is perform for each partition of different granularities separately and features of one partition does not affect features of the other ones. If one partition of a granularity contains the target concept, the result of classification on its features will be positive. Therefore, that image is labeled as positive (with the target concept) and that image is ranked in accordance with the best value of all the positive classifications of different granularities.

Furthermore, in our method, different granularities have the same worth in detection but in SPM bigger granularities have higher weights (for their features) which is not true since bigger granularities do not necessarily represent the concept better than the smaller granularities. In addition, it is observed that using "very" small granularities leads to the decrease of the performance. Furthermore, for both of the proposed and SPM methods, the K-NN classifier gives the better results than the SVM classifier. The reason is that the K-NN uses the multiplicative distance which is stable for the high-dimensional space (even for SIFT features). However, the SVM uses the radial basis function with "Euclidean distance" that its instability leads to the performance degradation.

Of course, the proposed method needs the manual detection of positive partitions in the training stage, which makes it labor-expensive for huge training datasets. If it is necessary, for solving this problem we can use semi-supervised algorithms or sampling techniques. On the other hand, the SPM is an unsupervised method and does not require human labor.

Table 1. Notations for different cases of fusions of different granularities.

Notation	Cases of fusion
fusion1	1×1 + 1×2
fusion2	1×1 + 2×1
fusion3	1×1 + 1×2 + 2×1
fusion4	1×1 + 1×2 + 2×1 + 2×2
fusion5	1×1 + 1×2 + 2×1 + 2×2 + 3×3
fusion6	1×1 + 1×2 + 2×1 + 2×2 + 3×3 + 4×4

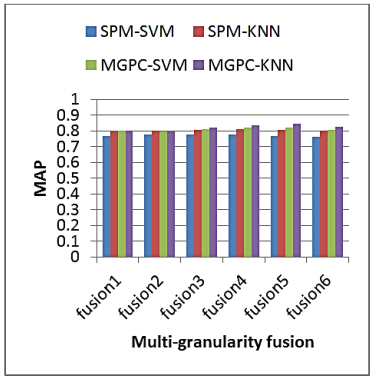


Fig. 5. MAP for different kinds of multi-granularity fusions for the proposed and SPM methods with SVM and K-NN classifiers for PASCAL VOC dataset.

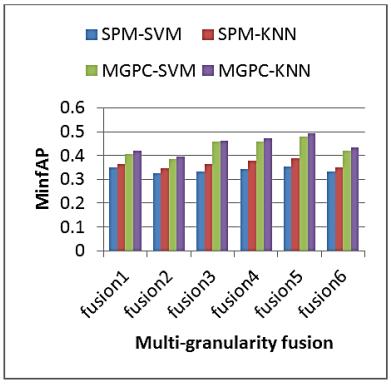


Fig. 6. MinfAP for different kinds of multi-granularity fusions for the proposed and SPM methods with SVM and K-NN classifiers for TRECVID dataset.

Now, we consider SVD features for different color spaces. For notational simplicity, the proposed method, which uses SVD features with multi-granularity partitioning and classification scheme, is referred to as SVDMGPC. Therefore, SVDMGPC-gray, SVDMGPC-RGB and SVDMGPC-HSV refer to the proposed method applied to images with 3 cases of gray-scale, RGB and HSV color spaces, respectively. In the experiments, the parameter *th* for the dimension reduction is set as 0.998 for all concept and all components of the color space. This value is obtained manually by subjective analysis for some values of *th*. However, it is possible that value of *th* is chosen adaptively for each concept using the objective quality metrics.

Figures 7 and 8 show the performance of the proposed method for different granularities individually for PASCAL VOC and TRECVID datasets. As shown in Figures 7 and 8, for both of datasets the vertical 1×2 grid has the best result among the granularities and 4×4 grid has the worst result. Moreover, the performance decreases for small granularities, i.e. 3×3 and 4×4 grids. This states that small granularities are not good choices for partitioning since small partitions may be not able to represent the target concept especially for non-object concepts (like in TRECVID dataset).

Figures 9 and 10 show the performance of the proposed method for different kinds of fusion between granularities for the PASCAL VOC and TRECVID datasets, respectively. Definition of fusions is as in Table 1. Based on Figures 9 and 10, with fusion of different granularities, performance usually increases. This shows the advantage of using multi-granularity partitioning and classification. When very small granularity (4×4 grid) is used for fusion, the performance decreases (especially for TRECVID dataset that has non-object concepts). This indicates that very small granularities cannot help for the detection even in the fusion.

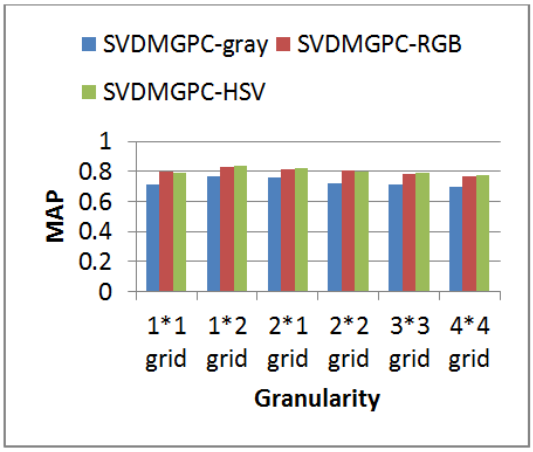


Fig. 7. MAP for different granularities for three cases of the proposed method for PASCAL VOC dataset.

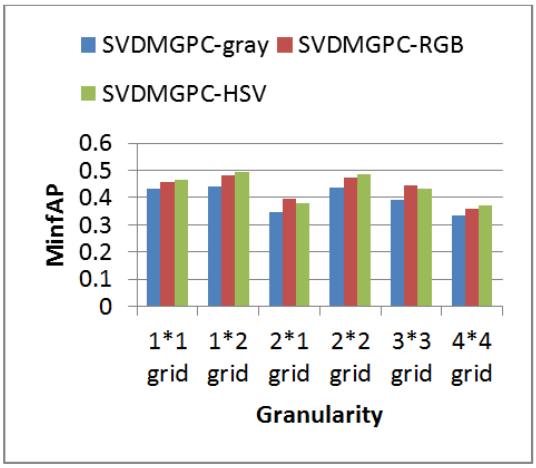


Fig. 8. MinfAP for different granularities for three cases of the proposed method for TRECVID dataset.

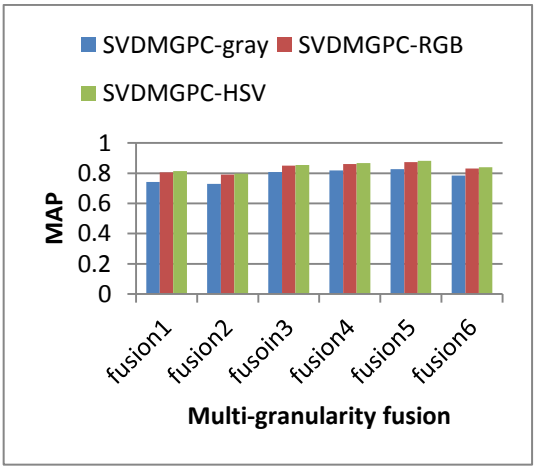


Fig. 9. MAP for different kinds of multi-granularity fusions for three cases of the proposed method for PASCAL VOC dataset.

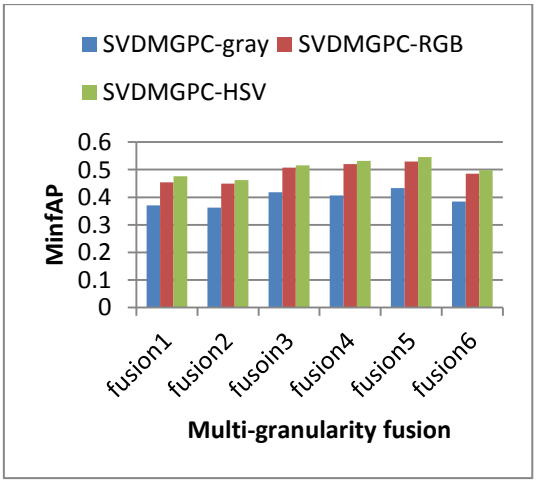


Fig. 10. MinfAP for different kinds of multi-granularity fusions for three cases of the proposed method for TRECVID dataset.

Now, we evaluate the proposed SVD features. For having a fair comparison in aspect of low-level features, widely-used local and global features in the literature, i.e. MPEG7 and BoW features, are selected for the comparison with the SVD features. Other referred works are not used for comparison. The reason is that either they

are variations of SIFT features (like Fisher vector) with their limitations or they do not consider just static low-level features (e.g. CNN performs both low-level feature and classification together). Additionally, we want to use multiplicative distance in our method since dimensionality of SVD features is high. Therefore, for a fair comparison we must use methods that multiplicative distance is applicable to them, and for this reason CNN cannot not be used.

For the comparison, MPEG-7 visual features [3] including 81-d color moments, 64-d color histogram, 62-d homogeneous texture, 80-d edge direction histogram, and the local feature [4] with 128-d SIFT descriptor are used. The compared method with these features is represented with CCWES. For fair comparison these features are extracted with the proposed multi-granularity partitioning scheme on HSV color space. For each feature of each partition, classification is performed separately using the K-NN algorithm with the multiplicative distance. Selection of the control power for each kind of the feature vector, determining positive test images, and forming ranked lists of images are carried out similar to our method. The best case of fusion, i.e. fusion5, is selected for the comparison of the global and local features with the proposed SVD features. Therefore, the difference between CCWES and SVDMGPC is just in the low-level features.

The performances of the proposed and CCWES systems for PASCAL VOC and TRECVID datasets are reported in Figures 11 and 12, respectively. Table 2 also reports the overall performance of the proposed and CCWES methods. Based on Table 2, using the proposed SVD features (in SVDMGPC-HSV and SVDMGPC-RGB) gives the superior performance over using the common and global features (in CCWES) for both of datasets. Moreover, as it can be seen from Figures 11 and 12, SVDMGPC-HSV and SVDMGPC-RGB have better performance than CCWES for detecting most of concepts. The reason is that in SVD feature texture, color and edge information is stimulatingly integrated with considering their relationship, and this information is sorted in accordance with their importance for representing concepts. But, these properties cannot be captured in the compared features.

Furthermore, based on Table 2, SVD features in the color space gives better performance than SVD features in the gray scale space. This shows that the color information can have an important impact on the concept detection performance. However, for some concepts in the two datasets the SVD features in the gray scale space gives better results than the SVD features in the color space. This indicates that color information does not always help for detection since some concepts may not be dependent on specific colors. One weakness of the proposed SVD features is that images should be in the same size. For solving this problem all images should be resized to the same size. Of course, this problem does not exist for the global and local features.

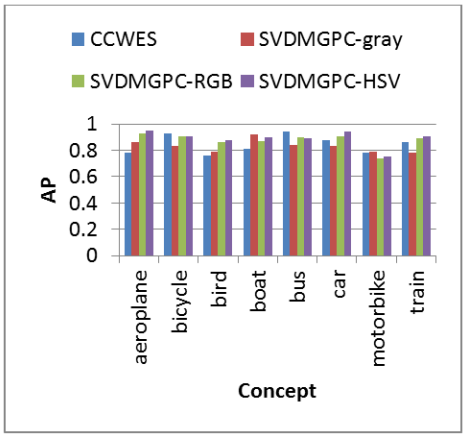


Fig. 11. AP for CCWES and three cases of the proposed methods for PASCAL VOC dataset.

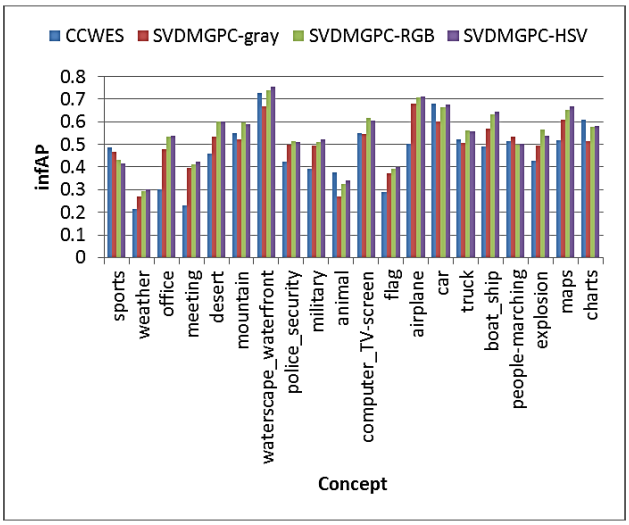


Fig. 12. infAP for CCWES and three cases of the proposed methods for TRECVID dataset.

Table 2. The overall performance of the CCWES and three cases of the proposed methods.

Methods	CCWES [3,4]	SVDMGPC-gray	SVDMGPC-RGB	SVDMGPC-HSV
MAP (PASCAL VOC)	0.8424	0.8263	0.8747	0.8820
MinfAP (TRECVID)	0.4637	0.4329	0.5297	0.5456

To confirm whether the improvement of the proposed method is statistically significant, we further conduct randomization test (suggested by TRECVID [32]) on the proposed and compared methods. In this test, the standard number of iterations in the randomization is 10000 and the standard level of significance is 0.05. The results of this test are shown in Table 3. P-value is the probability that the difference between two methods is due to chance. From these results, it is observed that for both of datasets the improvements of SVDMGPC-HSV and SVDMGPC-RGB over CCWES (and also SVDMGPC-gray) are statistically significant. Moreover, there is no significant difference between SVDMGPC-HSV and SVDMGPC-

RGB. Furthermore, the difference between SVDMGPC-gray and CCWES is not statistically significant.

Table 3. p-values of the significance test between methods for PASCAL VOC and TRECVID datasets.

Methods	p-value (PASCAL VOC)	p-value (TRECVID)
SVDMGPC-HSV vs. SVDMGPC-RGB	0.0706	0.1075
SVDMGPC-HSV vs. CCWES	0.0001	0.0332
SVDMGPC-HSV vs. SVDMGPC-gray	0.0005	0.0003
SVDMGPC-RGB vs. CCWES	0.0006	0.0397
SVDMGPC-RGB vs. SVDMGPC-gray	0.0014	0.0005
CCWES vs. SVDMGPC-gray	0.1976	0.4227

Total training and test time for the proposed and CCWES methods are shown in Table 4. The simulations have been carried out on a PC with Intel Core i7 CPU 2.79 GHz, and 8GB RAM with MATLAB. The proposed method consumes more training and test time. The reason is that extracting SVD features needs more time than extracting local and global features in CCWES. Moreover, SVD features have much more dimensionality than the features in CCWE and this leads to the more computations.

Table 4. Total training and test time (hours) for the proposed and compared methods for PASCAL VOC and TRECVID datasets.

Methods	CCWES [3,4]	SVDMGPC-gray	SVDMGPC-RGB	SVDMGPC-HSV
Training time (PASCAL VOC)	12.4	18.6	53.8	56.2
Test time (PASCAL VOC)	1.9	7.5	19.3	21.7
Training time (TRECVID)	51.6	75.0	212.7	216.4
Test time (TRECVID)	9.2	32.7	88.3	91.5

6. Conclusion

In this paper, new kind of static visual features, namely, right and left singular feature vectors and singular value feature vector, were proposed which were derived by applying SVD directly to the raw images. These features were different from features of methods in which SVD was applied to the low-level features of images (like in PCA or LAS techniques). Particularly, the proposed SVD features had this advantage that in which edge, color and texture information was integrated simultaneously and was sorted in accordance with their relationship and importance for the concept detection.

Additionally, feature extraction was performed in the multi-granularity manner. Furthermore, in the proposed method classification was carried out for each partition of each granularity separately, in contrast to the existing systems in which classification was performed for the whole image not for each partition. The proposed multi-granularity partitioning and classification had this advantage that the results of classifications on partitions with and without the target concept were not affected each other. This led to the performance improvement of the concept detection. Since usually feature vectors were high-dimensional even after the dimension reduction, classification was carried out by the K-NN algorithm with a new distance function, the multiplicative distance. This distance function was stable in the high-dimensional space, and was also usable in the low-dimensional space.

Experimental results showed the superiority of the multi-granularity partitioning and classification method over the spatial pyramid matching method and classification on the whole image and also the superiority of the proposed SVD features over the widely-used local and global features for the concept detection. However, the proposed method consumes more training and test time. The reason is that extracting SVD features needs more time than extracting local and global features in the compared method. Moreover, SVD features have much more dimensionality than the compared features, and this results in more computations.

References

[1] M. Jiu and H. Sahbi, “Nonlinear Deep Kernel Learning for Image Annotation,” IEEE Trans. Image Process., vol. 26, no. 4, pp. 1820-1832, Apr. 2017.

[2] H. Kaur and V. Dhir, “Local maximum edge cooccurance patterns for image indexing and retrieval,” Int’l Conf. Signal and information Processing, 2016.

[3] Moving Picture Expert Group. [Online]. Available: http://www.chiariglione.org/mpeg

[4] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, “Representations of keypoint-based semantic concept detection: A comprehensive study,” IEEE Trans. Multimedia, vol. 12, no. 1, pp. 42–53, Jan. 2010.

[5] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is nearest neighbors meaningful?” in Proc. Seventh Int’l Conf. Databasse Theory (ICDT ’99), 1999, vol. 1540, pp. 217–235.

[6] C.-M. Hsu and M.-S. Chen, “On the design and applicability of distance functions in high-dimensional data space,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 4, pp. 523–536, Apr. 2009.

[7] R. J. Durrant and A. Kaban, “When is ‘nearest neighbour’ meaningful: A converse theorem and implications,” J. of Complexity, vol. 25, no. 4, pp. 385–397, 2009.

[8] J. Mansouri and M. Khademi, "Multiplicative Distance: A method to alleviate distance instability for high-dimensional data," Knowledge and Information Systems, vol. 45, no. 3, pp. 783-805, 2015.

[9] Y, Han, Y. Yang, Y. Yang, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," IEEE Trans. Neural Networks and Learning Systems, vol. 26, no. 2, pp. 252–264, Feb. 2014,

[10] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," IEEE Trans. Pattern Anal. Machine Intell., vol. 34, no. 3, pp. 465–479, Mar. 2012.
[11] P. Srivastava and A. Khare, "Integration of wavelet transform, Local Binary Patterns and moments for content-based image retrieval," Journal of Visual Communication and Image Representation, vol. 42, pp. 78-103, 2017.
[12] X. Zhang and C. Liu, "Image understanding based on histogram of contrast," Signal, Image and Video Processing, vol. 10, no. 1, pp 103–112, 2016.
[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 2169-2178, 2006.
[14] J. Sanchez and J. Redolfi, "Exponential family Fisher vector for image classification," Pattern Recognition Letters, vol. 59, pp. 26-32, July 2015.
[15] K. Lim and H. Wang, "Sparse Coding Based Fisher Vector Using a Bayesian Approach," IEEE Signal Processing Letters, vol. 24, no. 1, pp. 91-95, Jan. 2017.
[16] Y. Shi, Y. Wan, K. Wu and X. Chen, "Non-negativity and locality constrained Laplacian sparse coding for image classification," Expert Systems with Applications, vol. 72, pp. 121-129, 2017.
[17] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-Modal Retrieval With CNN Visual Features: A New Baseline ," IEEE Trans. Cybernetics, Vol. 47, no. 2, pp. 449-460, Feb. 2017.
[18] G. Wang and Q. M. J. Wu, Advances in Pattern Recognition: Guide to Three Dimensional Structure and Motion Factorization, London: Springer, 2011.
[19] H. Yanai, K. Takeuchi, and Y. Takane, Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition, New York: Springer, 2011.
[20] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 347–364, Apr. 2012.
[21] M. Radovanovi'c, A. Nanopoulos, and M. Ivanovi´c, "On the existence of obstinate results in vector space models," in Proc. 33rd Int. ACM SIGIR conference on Research and development in information retrieval, New York, 2010, pp. 186–193.
[22] J. Hegde, "Time course of visual perception: Coarse-to-fine processing and beyond," Progress in Neurobiology, vol. 84, pp. 405–439, 2008.
[23] M.D. Menz and R.D. Freeman, "Stereoscopic depth processing in the visual cortex: A coarse-to-fine mechanism," Nat. Neurosci., vol. 6, pp. 59–65, 2003.
[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). The PASCAL Visual Object Classes Challenge Results [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[25] M. Sanderson, "Test collection based evaluation of information retrieval systems," Foundations Trends Inform. Retrieval, vol. 4, no. 4, pp. 247–375, 2010.
[26] P. Over, G.M. Awad, W. Keraaij, and A.F. Smeaton, "TRECVID 2007-overview," in TRECVid 2007 - Text REtrieval Conference TRECVid Workshop, Gaithersburg, Maryland, Nov. 2007.

Kamran Farajzadeh is Ph.D student of IT management in Islamic Azad University, Science and Research Branch, Tehran, Iran. His major research interests are wireless sensor networks, Internet of Things, image processing, robotic, medical informatics and IT management.

Esmail Zarezadeh was born in Tehran, Iran. He received the B.Sc degree in electrical engineering, and M.Sc degrees in electrical engineering and space engineering in 2008, 2013, respectively. Now he is Ph.D Student in electrical engineering at the Amir Kabir University of Technology (Tehran Polytechnic), Tehran, Iran. His working experiences are Radar systems, digital communication, application of sensing in multi antenna network and Adhoc systems. His research interests are MTM, RF/microwave circuits design and image processing.

Jafar Mansouri received Ph.D of electrical/communication engineering at Ferdowsi University of Mashhad, Iran, in 2015. His main research interests include image and video processing and analysis, multimedia information retrieval, high-dimensional data analysis, machine learning and computer vision.

Investigating the Effect of Functional and Flexible Information Systems on Supply Chain Operation: Iran Automotive Industry

Abbas Zareian*
Department of Information Technology Engineering, Mazandaran University of Science and Technology,Babol,Iran
abbaszareian92@gmail.com
Hamed Fazlollahtabar
Department of Technology, Mazandaran University of Science and Technology,Babol,Iran
hfazl@alumni.iust.ac.ir
Iraj Mahdavi
Department of Industrial Engineering, Mazandaran University of Science and Technology,Babol,Iran
irajash@rediffmail.com

Received: 16/Nov/2016 Revised: 01/Jan/2017 Accepted: 30/Aug/2017

Abstract

This research studies the relationship between supply chain and information system strategies, their effects on supply chain operation and functionality of an enterprise. Our research encompasses other ones because it uses a harmonic structure between information systems and supply chain strategies in order to improve supply chain functionality. The previous research focused on effects of information systems on modification of the relationship between supply chain strategies and supply chain function. We decide to evaluate direct effects of information systems on supply chain strategies. In this research, we show that information systems strategy to improve the relationship between supply chain and supply chain strategies will be. Therefore, it can be said that creating Alignment between informational system strategy and supply chain strategies finally result in improvement of supply chain functionality and company's operation.

Keywords: Functional Informational Systems; Flexible Informational Systems; Supply Chain Performance.

1. Introduction

In global competitions of this era, different products must be accessible to customers according to their demands. Demand of customer for high quality and fast servicing increased the pressures that didn't exist before. As a result, companies cannot do everything individually anymore. According to this, activities such as supply and demand programming, providing material, production and programming products, product maintenance services, inventory control, distribution, delivery and customer service, which were done in company level, moved to supply chain now [1]. Generally, supply chain is a chain which includes all the activities related to product flow and material conversion, from providing material to final delivery [2]. There are two other flows in product flow one of which is information flow and the other one is financial and credit source flow. While supply chain operation plays a great role in improving company's functionality and its final success, we always try to find a way for increasing supply chain functionality [3-7]. Many tools are used in researches and actions for this among which information technology and information systems were the most effectives. There raise two questions here:

1. Does using information system improve supply chain functionality?
2. What kind of relationship must exist among information systems and supply chain in order to increase effectiveness?

One of the aspects of successful management in supply chain lies in measuring and monitoring information of functional parameters and its key function. Therefore, it is important for the company to choose those systems which are in line with supply chain that is choose information systems that facilitate processes of supply chain and provide information about parameters which identify special aims of supply chain strategies. If this correlation don't exist between supply chain and information system strategies, this relationship will not only be ineffective but also results in losing organizational capital [8]. Therefore, suitable relationship between supply chain and information systems needs a basis for analyzing how information processing needs to different supply chains can be supported via information system programs. This research shows the conceptual space between alignment of different information system strategies and supply chain strategies providing theoretical and experimental basis for analyzing the benefits of information system programs for supply chain [9, 10]. We evaluate the moderated relationships between strategies of supply chain (special types of strategic aims and purposes that supply chain can have) and information system strategies (information of information system share for supply chain) and their effects on supply chain function (flexibility of supply chain, integration and answering customers) and companies functionality (how a company reaches its financial aims). With theorization about supply chain and information system literature in a

* Corresponding Author

framework via information process theory, we provide the hypothesis of our suggestion with positive effects of two strategies of information systems – functional and flexible – on relationships between lean and agile supply chain strategies. Our research question includes the following cases which can be answered during research:

1. Can supply chain strategies improve supply chain functionality?
2. Can functional information systems improve lean supply chain strategies?
3. Can flexible information systems improve agile supply chain system?
4. Can improvement of supply chain improve company's functionality at the end?

In this section, we stated _the general research and provided primary descriptions for the main elements of research and the necessity and importance of this research. Section 2, provides a review of literature associated with precise description of research elements and the background of research. Section 3, includes two parts: in the first part conceptual research model and research hypotheses will be stated and part two describes the research method. Section 4, analyzes the conducted case study in an automotive industry. Section 5, gives the conclusion and further research suggestions.

2. Review of Literature

In this section, reviews on the related works are categorized into supply chain and information systems and also the back ground is detailed.

2.1 Supply Chain

Supply chain includes all the related activities to flow and conversion of products from raw material step to delivery to final user and information flows of them. A remarkable notion in supply chain is supply chain strategy. Supply chain strategy reflects the "nature" of it and shows its specific aims and purposes [11]. Category of supply chain strategies show that it can focus on costs and weaknesses functionalities, flexibility and quick answer or a mixture of them. Different strategies of supply chain include noble, fast, flexible or a mixture of them that we describe only lean and agile in brief:

Lean supply chain: it is based on decrease of costs and flexibility, focusing on improvement of processes. Concentration of this supply chain is on decreasing waste products and increasing additional value and it aims at fulfillment of customers' needs and maintaining benefits.

Agile supply chain: its aim is to create answering ability with effective costs to unpredictable changes in market in terms of type and volume. One of the most important benefits of fast supply chain is customers' satisfaction, delivery speed, introducing new product, and decreasing delivery time [12].

2.2 Information Systems

An information system is a combination of integrated elements which support decision makings and organization controls via gathering, processing, saving and distributing informational data. This system helps to provide coordination in organizational operations and helps managers and employers in organizations to analyze and simulate organizational problems [13].

Information Systems Strategies

These strategies of a company develop the nature of information system strategies in share of functional programs of it and reveal itself in company status toward information systems [14]. Two strategies of functional and flexible strategies of information systems will be evaluated now:

Function Informational Systems

This strategy tends to functionality support inside and between organizations. In supply chain, information systems include automatic workflow, electronic exchange systems and connection process systems for functionality to monitor and control between and inside organizational processes. These systems facilitate operational functionality in supply chain via recording transactions, providing information about easy accessibility to them, structuring inside and between companies works according to standard activities and using standard protocols for simplifying information communication between them [15]. One of the most important types of it is organizational resource programming, data electronic exchange and automation.

Flexible information systems.It points at a type of information systems which focus on capability of company in order to market accessing and support in terms of rapid strategic decision making. Supply chain includes information system strategies that determine the share of programs that support market information systems and strategic decision making systems. The most important flexible information systems are communication with customer management and market analysis [16].

2.3 Research Background

Bendoly and Jacobs [17] considered logistic integration improvement as a factor for having operational benefits including decreasing costs, delaying dangers, improving selling, distribution of customer services and service levels and customer content. Bendoly and Schoenherr [16] evaluated creation of administrative structure by using information technology in integration of supply chain and customers and suppliers. This structure aims at using information technology in order to create a relationship for better understanding of customers' needs. Our research encompasses other ones because it uses a harmonic structure between information systems and supply chain strategies in order to improve supply chain functionality. The previous researches Okyere [18]

focused on effects of information systems on modification of the relationship between supply chain strategies and supply chain function.

We decide to evaluate direct effects of information systems on supply chain strategies.

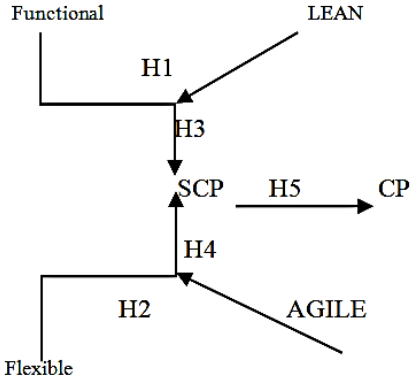
3. Research Model

In this section, the research specifications are given and the hypotheses and research questions are explained. The variables are extracted from supply chain and information systems.

3.1 Conceptual Research Model

We state that:

1. Functional information systems can improve the operation of supply chain lean strategy or empower it.
2. Flexible information systems can improve the functionality of agile supply chain strategy or empower it.
3. Improving the functionality of lean and agile strategies can improve supply chain functionality at the end.
4. Improved function of supply chain can improve total function of company.



3.2 Research Main Hypotheses

H1: Functional information system has significant effect on lean supply chain strategy.

H2: Flexibility has significant effect on agile supply chain.

H3: Lean strategy has significant effect on supply chain function.

H4: Agile strategy has significant effect on supply chain function.

H5: There is significant relationship between supply chain function and company function.

3.3 Research Method

We have three steps for this research:

1. Providing questionnaire and primary studies
2. Data collection in high level
3. Data analysis

Step one: We need questionnaires which can evaluate research hypotheses precisely. Therefore, questionnaires must be valid. Validity deals with the fact that how much measuring tools provide equal results in equal situation.

After evaluation of questionnaire,

the averages of hit ratio: 98%,
agreement between judges: 95%,
Cohen's kappa coefficient: 95%.

This shows excellent level of judgment agreement [19]. The questionnaire is as follows:

It has 6 parts:

Part one: lean supply chain functionality and operation questions

Standard product production amount	LS1
Decrease of trashes and wastes	LS2
Investment management via sending list of demands	LS3
Frequent inspection of products	LS4
Management with product quality according to needs	LS5

Part two: agile supply chain effect evaluation

Effective answer to design needs	AS1
Rapid answer to orders of customers	AS2
Ability to controlling changes in product design	AS3
Maintaining more than market need capacity for rapid answering	AS4
Making products custom-made via adding special models	AS5

Part three: effects of functional information decisions

Improving operational efficiency between suppliers and company	EIS1
Investment management between suppliers and company	EIS 2
Raw material planning management	EIS 3
Production management between suppliers and company	EIS 4
Coordination between functional (production and information) for supplier and production line	EIS 5

Part four: ability of effective information systems evaluation

Introducing new products and services of market	FIS1
Monitoring situations and market changes	FIS 2
Answering market changes	FIS 3
Changing product design	FIS 4
Ability in communication with customers	FIS 5

Part five: determination of supply chain function

Ability to investigation non-standard order	SCP1
Ability to answer special demands of customers	SCP2
Ability to produce goods with different features	SCP3
Ability to setting speed in answering changes in customer's demands	SCP4
Ability to accelerate introducing improved products	SCP5
Ability to introduce new products	SCP6
Quick answer to customers	SCP7
Common activities of company and business partners	SCP8
Improving integration level via informational system	SCP9
Shortening time of order - delivery	SCP10

Part six: company function

Market share	CP1
Investment return	CP2
Stock growth	CP3
Benefit and profit differences	CP4
General competitive situation of company	CP5

Range 1 (weak) to 5 (excellent) is used in the questionnaire.

Step two: Analyzing unit is in central company. Senior executive managers are chosen from selling/ production/ supply chain sections. This research evaluates

companies with advanced information systems. Our case study is Iran Best Automotive Industry.

Step three: after answering questionnaires via senior managements, the answers will be evaluated with modeling structural equation analysis software. Then t-value will evaluate the strength and meaningfulness of research hypotheses to show meaningful hypotheses.

4. Statistical Analysis

Data analysis is a multiple step process which use gathering tools in sample to summarize, code, categorize, etc. and finally processing data to analyze them. Data will be processed conceptually and experimentally here and different statistical techniques play an important role here.

This section analyses data collected via questionnaires using statistical suitable techniques and results will be provided by descriptive statistical techniques. Statistical indexes are used including frequency, frequency percentage, and cumulative frequency percentage for information analysis. Hypotheses are tested via modeling technique of structural equations.

4.1 Answerers' Characteristics

Statistical descriptive indexes are used for describing general features. Frequency of answerers is evaluated in terms of age, education, gender and work experience and related charts are provided.

Gender

169 people, more than 72% of the answerers were men. 65 people, more than 27% were women.

Age

24 people were less than 30. 72 people were 30-40 and 87 people are 41-50 and 51 people are more than 50.

Education

16 people have high school degree or less, 55 people have associate degree, 117 people have BA which has the most frequency and 46 people have MA and higher degrees.

Work Experience

26 people have less than 5 years' work experience. 55 people have 5015 years, 84 people have 16-25 years and 69 people have more than 25 years' work experience.

4.2 Descriptive Statistics of Research Variables

As described, in descriptive methods it is tried to provide tables and use descriptive statistic tools including central and dispersion indexes to describe research data to clear the subject. The following table includes descriptive statistics for all the research variables. In the first part, the most important central and dispersion indexes are provided. Among central indexes, average, median and among dispersion indexes standard derivation are used. All the variables are provided maximum and minimum and difference of these numbers provides one of the simplest dispersion indexes; that is changing ratio. SPSS software calculated the element of this table:

Table 1. Descriptive statistics of Research variables

Variable	Minimum	Maximum	Average	Variance
Functional information system	1.60	3.56	5.00	0.41
lean strategy	1.60	3.60	5.00	0.45
Agile strategy	1.60	3.57	5.00	0.56
Flexible information system	1.40	3.68	5.00	0.50
Supply chain function	1.50	3.62	5.00	0.42
Company function	1.60	3.74	5.00	0.43

Note that the reliability of the questionnaire is computed by Cronbach's alpha equal to 0.81 which is desirable.

4.3 Testing Normalness of Data Distribution

Kolmogorov-Smirnov technique is used in this research for determining normalness of data distribution. In accepting data analysis and modeling of structural equations there is no need to normalness of all data yet factors must be normal. Therefore, considering data normalness in 0.05 meaning fulness level is tested via Kolmogorov- Smirnov technique. The following hypothesis must be set for this test:

H0: data distribution of all variables are normal.

H1: data distribution of all variables are not normal.

Normalness test result of data is provided in Table 2.

As seen in Table 2, meaningfulness is higher than 0.05 in all cases. So there is no reason to deny H0. It means that data distribution is normal and parametrical tests can be done.

Table 2. Data distribution normalness test

Variable	Freedom degree	Amount of K.S	status
Functional information system	234	0.08	normal
Lean strategy	234	0.08	normal
Agile strategy	234	0.08	normal
Flexible information system	234	0.08	normal
Supply chain operation	234	0.07	normal
Company operation	234	0.09	normal

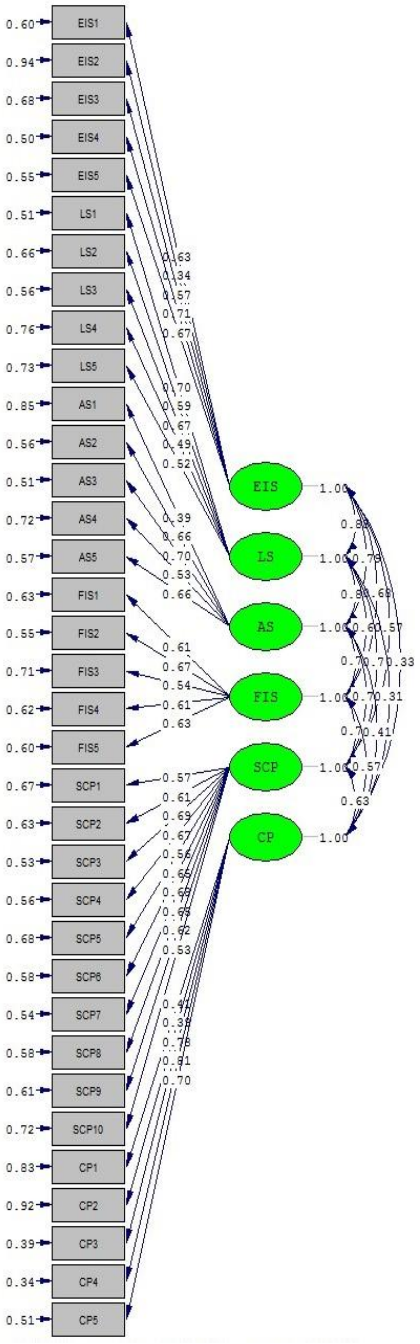
4.4 Factorial Acceptance Analysis of Research Questionnaire

This research uses questionnaire for gathering data. Therefore, factorial acceptance analysis evaluated general structure of research in terms of validity. For factorial acceptance analysis and modeling structural equations, t statistic and standard load factor are measured. The following regulation is true generally:

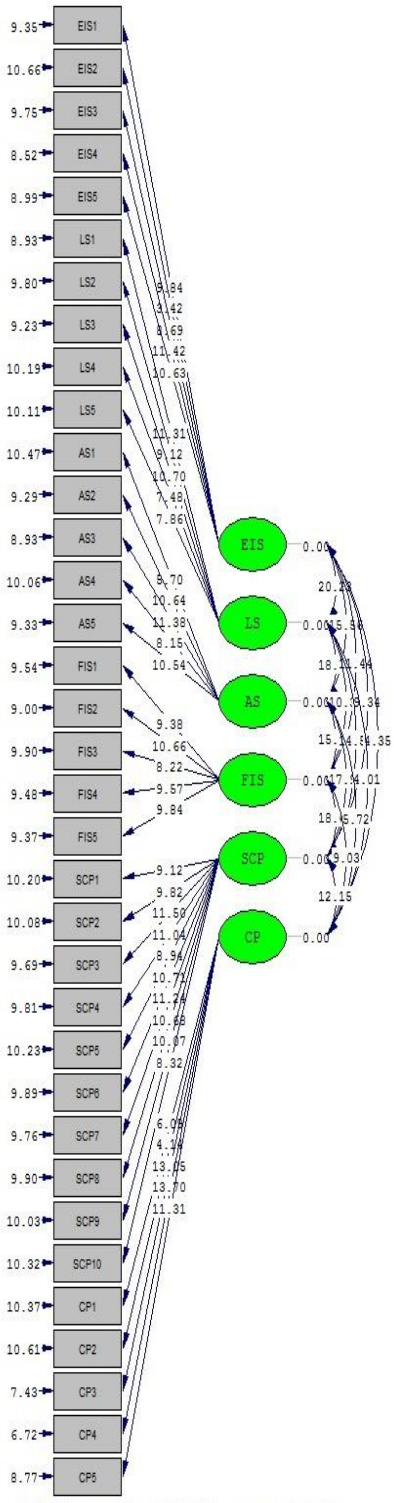
The relationship power between factor (hidden variable) and obvious one is shown via load factor. Load factor is a number between 0 and 1. If it is less than 0.3, the relationship will be considered as weak and will be taken for granted. Load factors between 0.3 and 0.6 are acceptable and if it is more than 0.6 it can be considered as favorable. After determining correlation between variables, meaningfulness test must be done. In order to evaluate meaningfulness of relationships t test and t value

is used. Because meaningfulness is studied in 0.5 error level, if loads of t-value is less than 1.96, the relationship is not meaningful and it will be shown with red color in LISREL software.

Factorial acceptance analysis is shown in figure ١. Standard load factor accepts power of relationship measurement between each factor with its obvious variable is more than 0.3 in all cases. Therefore, this questionnaire is acceptable. After measuring standard load factor, meaningfulness must be tested. According to the results in figure ٢, t load factor of each aspect in 0.05 confident levels is more than 1.96. Therefore, the correlations are meaningful.



Chi-Square=1203.54, df=545, P-value=0.06412, RMSEA=0.024
Fig. 1. Standard load factor of questionnaire



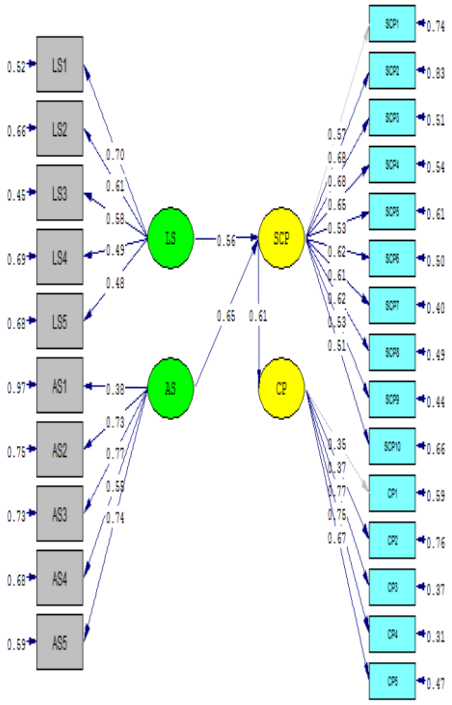
Chi-Square=1203.54, df=545, P-value=0.06412, RMSEA=0.024
Fig. 2. t-value of questionnaire

4.5 Final Model of Relationships between Variables and Evaluation of Research Hypotheses

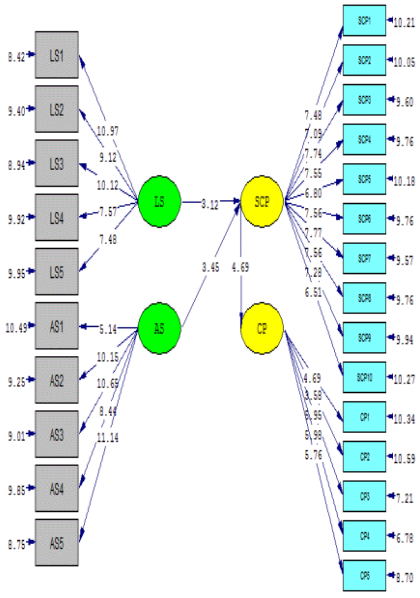
Final structural equation model for measuring the relationship between main factors of research is used. Because each factor includes some hidden variables, the average answer of each variable is measured and that variable is used in the final model as an obvious one.

Final model is provided in figure 4. This model is drawn using LISREL software. The results from data meaningfulness measurement are shown in figure 5.

The impact of the strategy of lean and agile on supply chain performance without information systems are shown in Figures 3 and 4, respectively.



Chi-Square=726.85, df=271, P-value=0.00000, RMSEA=0.035
Fig. 3. Results of accepting final model of relationship between research variable



Chi-Square=726.85, df=271, P-value=0.00000, RMSEA=0.035
Fig. 4. Results of accepting final model of relationship between t-value

In order to fit structural research model a number of goodness of fix indexes are used. One of the general indexes for measuring free parameters in calculating fitness indexes is Chi do index which is calculated via dividing *Chi* do to freedom degree. If it is between 1 to 5, the amount is favorable. $\frac{\chi^2}{df} = \frac{726.85}{271} = 2.68$

In order to determine fitness of model some of goodness fit indexes are used which are shown in Table 3. Because RMSEA is less than 0.1, model fitness is good. Other goodness of fit indexes is also acceptable.

Table 3. *Goodness* of Fit *Index*

Fitness index	SRMR	RMSEA	GFI	AGFI	NFI	NNFI	IFI
Acceptable amounts	<0.1	<0.1	>0.9	>0.9	>0.9	>0.9	0-1
Calculated amounts	0.035	0.020	0.92	0.95	0.96	0.94	0.98

Lean supply chain strategy has correlation with function of supply chain performance.

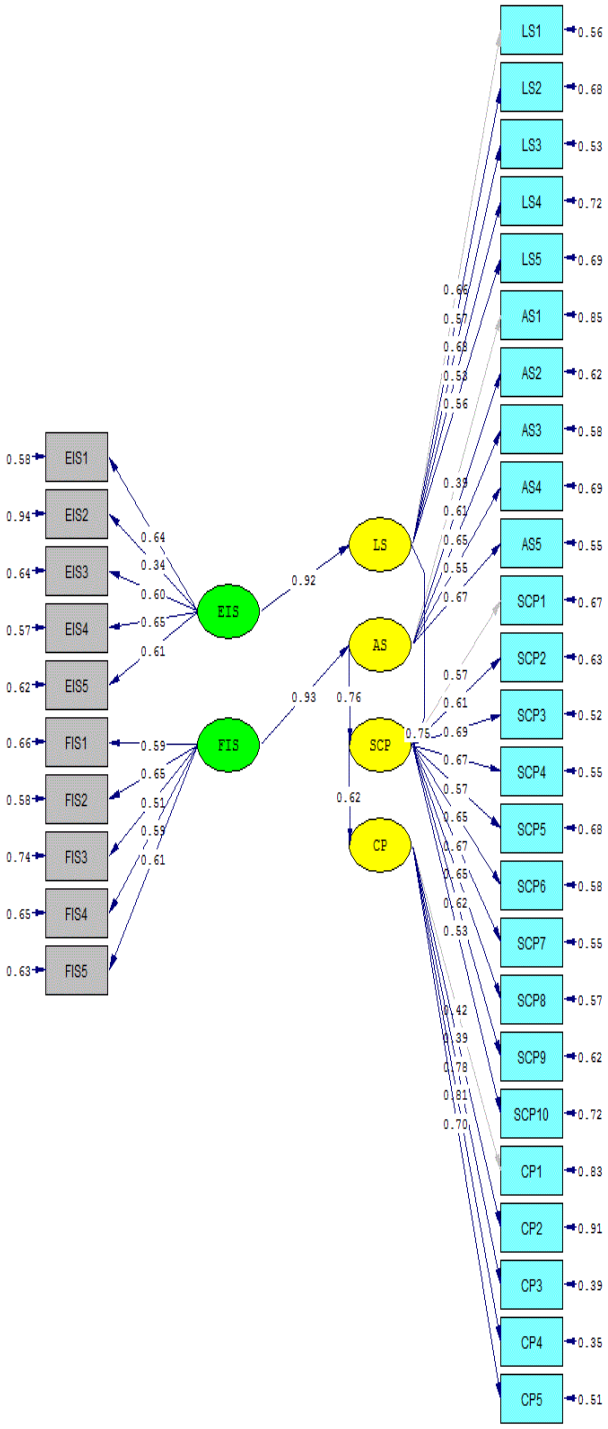
Lean supply chain strategy has correlation with higher function levels of supply chain. According to the calculations, standard load factor of noble supply chain's structure and supply chain strategy is 0.56 that shows there is a favorable and strong relationship between these two. T load factor is 2.74 which show that the correlation is meaningful. Therefore hypothesis 3 is accepted; it means that noble supply chain strategy has correlation with higher function levels of supply chain.

Agile supply chain strategy has correlation with higher level of supply chain performance.

Agile supply chain strategy has correlation with higher level of supply chain function. According to the calculations, standard load factor of fast supply chain's structure and supply chain is 0.51 that shows there is a favorable and strong relationship between these two. T load factor is 2.21 which show that the correlation is meaningful. Therefore hypothesis is accepted; it means that fast supply chain strategy has correlation with higher level of supply chain function.

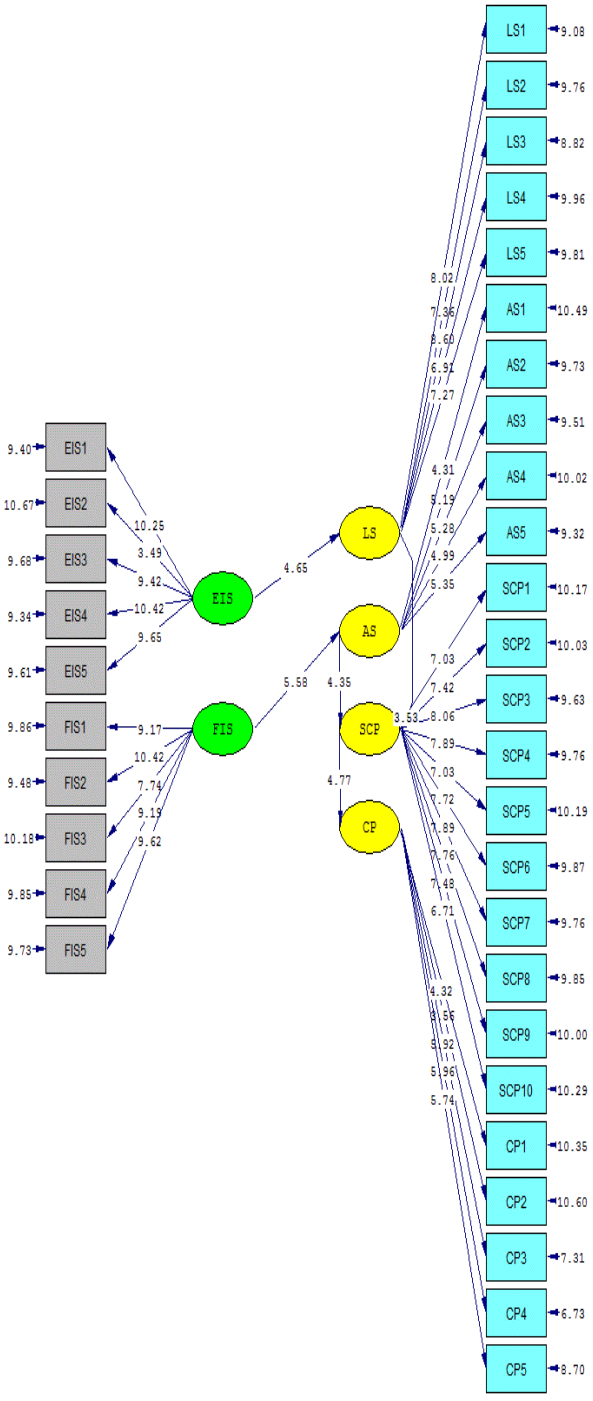
Improved function of supply chain improves total function of company.

Improved function of supply chain improves total function of company. According to the calculations, standard load factor of supply chain's structure and company's function is 0.61 that shows there is a favorable and strong relationship between these two. *T* load factor is 4.69 which show that the correlation is meaningful. Therefore hypothesis is accepted; it means that improved function of supply chain improves total function of company.



Chi-Square=1279.16, df=554, P-value=0.02051, RMSEA=0.045

Fig. 5. Results of accepting final model of relationship between research variables



Chi-Square=1279.16, df=554, P-value=0.02051, RMSEA=0.045

Fig. 6. Results of accepting final model of relationship between t-value

In order to fit structural research model a number of goodness of fix indexes are used. One of the general indexes for measuring free parameters in calculating fitness indexes is Chi square index which is calculated via dividing Chi do to freedom degree. If it is between 1 to 5, the amount is favorable.

$$\frac{\chi^2}{df} = \frac{1279.16}{554} = 2.309$$

In order to determine fitness of model some of goodness fit indexes are used which are shown in Table 4. Because RMSEA is less than 0.1, model fitness is good. Other goodness of fit indexes is also acceptable.

Table 4. *Goodness* of Fit *Index*

Fitness index	SRMR	RMSEA	GFI	AGFI	NFI	NNFI	IFI
Acceptable amounts	<0.1	<0.1	>0.9	>0.9	>0.9	>0.9	0-1
Calculated amounts	0.036	0.045	0.93	0.97	0.94	0.95	0.96

H1: Information systems improve lean supply chain strategy for functionality.

According to the calculations, standard load factor of functional information systems' structure and lean supply chain strategy is 0.92 that shows there is a favorable and strong relationship between these two. T load factor is 4.65 which show that the correlation is meaningful. Therefore hypothesis 1 is accepted; it means that information systems improve lean supply chain strategy for functionality.

H2: Flexible information system improves agile supply chain strategy.

According to the calculations, standard load factor of flexible information systems' structure and agile supply chain strategy is 0.93 that shows there is a favorable and strong relationship between these two. T load factor is 5.58 which show that the correlation is meaningful. Therefore hypothesis 2 is accepted; it means that flexible information system improves agile supply chain strategy.

H3: Lean supply chain strategy has correlation with function of supply chain performance.

According to the calculations, standard load factor of lean supply chain's structure and supply chain strategy is 0.56 that shows there is a favorable and strong relationship between these two. T load factor is 2.74 which show that the correlation is meaningful. Therefore hypothesis 3 is accepted; it means that lean supply chain strategy has correlation with higher function levels of supply chain.

H4: Agile supply chain strategy has correlation with higher level of supply chain performance.

According to the calculations, standard load factor of agile supply chain's structure and supply chain is 0.51 that shows there is a favorable and strong relationship between these two. T load factor is 2.21 which show that the correlation is meaningful. Therefore hypothesis 4 is accepted; it means that agile supply chain strategy has correlation with higher level of supply chain function.

H5: Improved function of supply chain improves total function of company.

According to the calculations, standard load factor of supply chain's structure and company's function is 0.62 that shows there is a favorable and strong relationship between these two. T load factor is 4.77 which show that the correlation is meaningful. Therefore hypothesis 5 is accepted; it means that improved function of supply chain improves total function of company.

5. Conclusion

In this paper, the impact of information systems on supply chain strategies and performance were investigated. One of the significant outcomes of the research was that strengthening information systems lead to improve the relationship between supply chain and supply chain strategies. Therefore, it can be said that creating balance between informational system strategy and supply chain strategies finally result in improvement of supply chain functionality and company's operation.

Further Research Suggestions:

1. According to the fact that green strategy of supply chain is of great importance today, evaluation of the most suitable information system strategy for this strategy and effectiveness of information systems on them can be good.
2. Information system strategies have the ability to create two by to and multiple links. In case of link between these strategies, how will the balanced information systems act?

References

[1] A.H. Safaei Ghadikolaee, et al. comparative Evaluation of lean, agile and lean and agile Supply Chain Strategies, Executive Management Bulletin, 2011.

[2] B. Shahabi. "Human Aspect of Organizational Fastness", Measure Sentence, no 175, Industrial Management Organization, 2006.

[3] g. Stevens. Integrating the supply chains, International Journal of physical distribution and material management, 1989.

[4] H.L. Lee. Aligning supply chain strategies California management, 2002.

[5] M.A. Vonderembse, M. Uppal, S.H. Huang, J.P. pismukes. Designing supply chains: towards theory development. International journal of production E Conomics, 2006.

[6] P. Hines, M. Holweg, N. Rich. Learning to evolve: A review of temporary lean thinking. International journal of operations & production management,2004.

[7] A. Agrawal, R. Shankar, M.K. Tiwari. "modeling the metrics of lean, agile and Lean agile supply chain An ANP-based approach", European journal of operational Research,2006.

[8] M. christopher and C.Ruther ford. "Creating supply chain Resilience through agile six sigma," critical Eye, 2004.

[9] G. Gunasekaran and E. Tirtiroglu. "Performance measures and metrics in a supply chain environment", 2001.

[10] A. Gunase karan, C. Patel, and R. Mc Gaughey. "A frame work for supply chain performance measurement", 2004.

[11] D. Lambert and L. Pohlen. "supply metrics". The International journal of logistics management, 2001.

[12] C.R. Wu, C.W. Chang, H.L. Lin. «A Fuzzy ANP based Approach to Evaluate Medical Organizational Performance», Information and Management Sciences, Vol.19 No. 1, pp. 53-74,2009.
[13] J. Sanderson, A. Cox. «The challenges of supply strategy selection in a project environment: evidence from UK naval shipbuilding » International Journal of Supply Chain Management, 2008.
[14] D. Kisperska Moron, A. Swierczek. «The agile capabilities of Polish companies in the supply chain: An empirical study», International Journal of Production Economics, 118, 217–224, 2009.
[15] S. Qrunfleh, M. tarafdar. supply chain information system strategy :impact on supply chain performance and firm performance Int. J. Production Economics, 2006.
[16] E. Bendoly, T. Schoenherr. ERP system and implementation benefits: implications for B2B E-procurement. International Journal of Operations and Production Management 25 (4), 304–319, 2005.
[17] E. Bendoly, F. Jacobs. ERP architectural/ operational alignment for order processing performance. International Journal of Operations & Production Management 24 (1), 99–117, 2004.
[18] S. Okyere, G. tuo. supply chain information system flexibility and performance. Int J Production Econom, 2014.
[19] Y. Qi, X. Zhao, C. Sheu. The impact of competitive strategy and supply chain strategy on business performance: the role of environmental uncertainty, 2011.

Abbas Zareian has been graduated in M.Sc. of Information Technology Engineering from Mazandaran University of Science and Technology, Babol, Iran. He works on enterprise resource planning and supply chain management considering information systems performance. He presented his papers in international conferences and journals.

Hamed Fazlollahtabar earned a BS.c and an MS.c in industrial engineering from Mazandaran University of Science and Technology, Babol, Iran, in 2008 and 2010, respectively. He received his Ph.D. in industrial and systems engineering from Iran University of Science and Technology, Tehran, Iran, in 2015. He recently completed a postdoctoral research fellowship at Sharif University of Technology, Tehran, Iran, in the area of reliability engineering for complex systems. He is on the editorial boards of several journals and on the technical committees of several conferences. His research interests are in robot path planning, reliability engineering, supply chain planning, and business intelligence and analytics. He has published more than 230 research papers in international books, journals, and conferences. He has also published five books out of which three are internationally distributed to academicians.

Iraj Mahdavi is the full Professor of Industrial Engineering at Mazandaran University of Science and Technology, Babol, Iran. He received his Ph.D. from India in Production Engineering. He is also in the editorial board of five journals and scientific committee member of international conferences. He was awarded as the best researcher of engineering area in Iran and among the best professors of Iran. He has published over 300 research papers. His research interests include cellular manufacturing, production planning, supply chain, fuzzy networks, digital management and intelligent operations management.

De-lurking in Online Communities Using Repost Behavior Prediction Method

Omid Reza Bolouki Speily*
Department of Information Technology & Computer Engineering, Urmia University of Technology, Urmia, Iran
speily@uut.ac.ir

Received: 10/Sep/2016 Revised: 01/Sep/2017 Accepted: 05/Sep/2017

Abstract

Nowadays, with the advent of social networks, a big change has occurred in the structure of web-based services. Online community (OC) enable their users to access different type of Information, through the internet based structure anywhere any time. OC services are among the strategies used for production and repost of information by users interested in a specific area. In this respect, users become members in a particular domain at will and begin posting. Considering the networking structure, one of the major challenges these groups face is the lack of reposting behavior. Most users of these systems take up a lurking position toward the posts in the forum. De-lurking is a type of social media behavior where a user breaks an "online silence" or habit of passive thread viewing to engage in a virtual conversation. One of the proposed ways to improve De-Lurking is the selection and display of influential posts for each individual. Influential posts are so selected as to be more likely reposted by users based on each user's interests, knowledge and characteristics. The present article intends to introduce a new method for selecting k influential posts to ensure increased repost of information. In terms of participation in OCs, users are divided into two groups of posters and lurkers. Some solutions are proposed to encourage lurking users to participate in reposting the contents. Based on actual data from Twitter and actual blogs with respect to reposts, the assessments indicate the effectiveness of the proposed method.

Keywords: De-Lurking; Post Similarity; Lurker; Online Community.

1. Introduction

Online community (OC)[1] refers to a group of users who have a common interest in a particular subject area, produce, and share online knowledge and information [1] - [3]. Most of these communities transfer their knowledge in the form of texts, hypertext and multimedia on Web Platforms or forums [1]-[3]. With the development of such forums as yahoo answers, Twitter or stack overflow, etc., the fundamental challenge is increased information sharing in these groups [4]. OC's are classified as direct and indirect communications [5]. Indirect communication is usually done through the establishment of two-way friendship and direct communication between two users is made by following other users for the latest posts created and not necessarily is this type of communication two-way. Despite the diversity in implementation, all these groups follow the same process: a user posts a message and the other user reposts it if he/she likes it. These posts are virally shared on the net. Accordingly, a certain post is made available to a large number of users [6]. In the present study, considering the nature of OCs, the communication among users is considered to be direct. Experts from different subfields have gathered together in these groups and the users follow the posts made by other users in shared areas according to their interests [3]. Nevertheless, there are many users who do not participate actively in these online communities. According to previous studies, users are divided into two groups: lurkers (nonparticipants) and posters (participants) [4]. In various papers, different definitions of lurkers have been provided [5]-[7]. Emphasizing a characteristic, each of these definitions has described such users. Generally, these users join an online community or group consciously; however, they do not post anything. Creating a post in online communities indicates users' participations. Regarding the level of this participation, no clear definitions have been provided in different references. In many references, the users who do not post anything in online communities are known as lurkers [5]; however, a time limit has been made for this nonparticipation in some other references [8]. For instance, if a user did not (re) post a post in the past year, s/he is considered a lurker. In addition, there are many disagreements over the effects of such users [4], [8], [9]. Some references have interpreted them as free riders who use sources free without providing communities with any benefits [10].

In the present study, we are looking for a method to select k influential posts for each user to be displayed when signing in. The idea has been addressed in several papers under titles such as post ranking, post refining, etc.[7] - [9]. The issue of selecting k posts for display does not mean to simply maximize the number of posts reposted but, in fact, its concern is to De-lurking in OCs.

1. The online community considered in this article has a similar structure to Twitter. In such online communities it is possible to track other users as well as view, comment, and re-post their posts.

* Corresponding Author

Delurking is a type of social media behavior in which a user breaks an "online silence" or habit of viewing an inactive thread to interact with a virtual conversation. The term means that the user typically does not participate in social media or online social activities [11].

Those posts are selected to be displayed that are more likely to be reposted by sub-users and cause de-lurking. To put it more formally, the purpose is to display for the user those posts that maximize a tree of reposts of which this user is the root.

In the proposed method, in addition to the structural characteristics of users on the network, the attention is paid to the attitude of reposting among users. Users are divided into two categories of posters and lurkers based on their reposting attitude. Lurkers are those who become a member in an online community but do not post (or simply post very little), are only readers, and are not active. Considering this type of users, the present study intends to reduce pure lurking given the time the important posts are available online, that is, the posts, which are more likely to be reposted by users, are displayed longer for lurkers than for posters.

In the second section of this article, the literature on the subject is reviewed. In the third section, the proposed method is presented according to the nature of lurkers. The fourth section contains a detailed evaluation of the proposed method and the fifth section is devoted to summing up the study done.

2. Literature Review

On the whole, numerous articles have been presented to solve the problem of reposting based on the identification of influential individuals in the network. In this section, we try to review the literature on the subject, especially the recent one. The article [12] evaluates the transfer of posts among bloggers. By analyzing timestamps on each post and transferring them on the network of Weblog users, it suggests that this sharing follows a specific model referred to in this article as waterfall model. The identification of these models is of great importance in the study on the sharing or transferal of information in (online and offline) communities. It is important to identify these models. The study [13] on blogsphere, addresses “epidemic” interests among different blogs with regard to the content cited or copied from another blog. By studying the cases, it estimates the relationship between two similar blogs. By relationship, it means the use of another post in the form of citation or copy. Another important point addressed in the study is the evaluation of the influence of a blog on another blog via a post. [14] is a study on sharing small pieces of text (for example, in connection with the news) used in other articles and other texts. For this purpose, a method has been implemented by which the source of each piece can be specified in the network. This makes it possible to study the structure of sharing in the network. The study aims to find the sources by which a post or posts are influenced.

Major constraints faced in the literature are, however, the absence of detailed knowledge on the network structure in which sharing takes place. In more recent researches, attempt has been made to obtain certain information about the network in which sharing takes place in addition to the collection of the data on sharing. For example, in [15], certain studies have been done on sharing tree in Facebook fan pages. In [16], Bakhshi conducted a study on sharing information in the online social game “second life”. In this article, by tracking the sharing of "gesture" an information unit in the game (which can be copied by other users), he obtained interesting results. Using simulation, Bakhshi showed that the possibility of transferring this information between two friends is more than that between two users who do not know each other. It has also been shown that certain users play a more important role in sharing the game information, called "adaptive users" in this study. Most of the methods presented on this subject are based upon the identification of individuals in the group influentially involved in sharing information. In this respect, the posts are mostly displayed to influential users so that, on reposting, more users may view these posts. These methods are mainly based on individuals’ statistical characteristics such as the number of follower users and number of reposts [10]. It is proven in [11] that there is an insignificant relationship between popularity (in terms of number of followers) and influence. Studies such as [20] and [21] have proposed more effective methods in terms of scalability and run-time compared to the ambitious method of Camp. In [22], a study has been conducted on a subset of Twitter data and a method has been presented for evaluating the influence of individuals regarding the characteristics in each issue. In [23], various criteria used to determine the influence of users have been investigated. Accordingly, this article has carried out a comprehensive study to explain the accuracy of the criteria such as indegree, repost and mentioning. The first criterion is the number of users following the subjects of the intended user. Repost refers to the number of times users repost the issues raised by the relevant users. Mentioning refers to the number of times that users mention the relevant user’s name. According to the experiments conducted in this study, having even more than a million followers does not guarantee a user’s influence. Given the existing studies and the nature of OCs, the present article tries to present a method for displaying the best information posts online needed by users leading to increased participation of the users. In this method, the importance of the posts displayed is taken into account in terms of post subject, user interest and information level in addition to the type of user (active or lurking).

3. *k* Influential Posts Selection

This section explains selection of k influential posts in OCs and discusses its characteristics.

3.1 Statement of Problem

Suppose that the OC is implemented under a social network. These groups are typically displayed as graphs of followers-followees. Users follow other users considering their interests and expertise needed. This directed graph is defined as G = (V, E) where E represents the relationship between users and V represents users. $(u,v) \in E$ shows that the user u follows the user v. If P represents the total of posts created in the whole online community, then an online social event (post) occurs when user u, creates post p at $t \in T$ time, represented as $post(u,p,t)$. In the same manner, when the user v shares the post p through the user u at time t, "reposting" occurs, represented as $repost(v,u,p,t)$. According to the definitions provided, the probability of repost can be defined as a function of the probability of repost $p : P \times V \times V \times T \rightarrow [1,0]$. In this function, T is the temporal domain. The probable repost of P posts by any user from the V users group, within the temporal domain T, includes the values between zero and one.

If σ is taken as the selection procedure of k influential post from among the candid posts (v, t) for the user v at any t time, the output $\sigma(v,t)$ is the k influential post to the user v at the time t. the total candidate posts for the user v are those already created or reposted the users of group V followed by the user v. Equation (1) shows the initial set of candidate posts for $t' < t$.

$$init(v,t) = \{p \in P \mid post(u,p,t') \cap (u,v) \in V\} \quad (1)$$

From this series of posts, the duplicate posts already displayed for the user in the previous time $t' < t$ should be removed. The candidate posts, as shown in equation (2), are as follows:

$$candid(v,t) = init(v,t) - \{p \in P \mid post(v,p,t') \cup \sigma(v,t')\} \quad (2)$$

Here, a formal definition is given for selection of the k post.

Definition (1): (selection of k influential posts) consider the graph online community as a graph G = (V, E) where V is the total of users and E is the total of follow-up communications among users. $k \in \mathrm{N}$ is an input issue and represents the number of influential posts selected from among candidate posts for display. This requires the definition of the procedure $\sigma : V \times T \rightarrow 2^P$ selecting for each user $v \in V$, and each timestamp t ($t \in T$), a k series of influential posts as $\sigma(v,t) \subseteq cand(v,t)$ where $|\sigma(v,t)| = k$ for display to the user v. If the number of posts is considered as |P|, there is $2^{|P|}$ different modes (sharing or not sharing any posts) for selection. The selection should be in a way that the number of network reposts depending on the model number of posts is maximized regarding the model of post sharing (section 2.3.3). Eq. (3) gives the mathematical expression for selection of influential posts.

$$Max \sum_{p \in P} |\{w \in V \mid \exists t \in T : post(w,p,t)\}| \quad (3)$$

3.2 Heuristic Method For Selection Of K Post

We provide an heuristic method for selection of k influential posts considering the computational capabilities and simplicity. This procedure is designed so as to be operational in online environments. In this section, the proposed method is introduced.

A- Similarity Between Posts (Post-Post)

For this purpose, two methods of topic similarity and geographical similarity are applied. For topic similarity, the posts of a user are considered as a set of words. According to the definition, this candidate post t_{wi} is topically similar to the posts of user v if it is related to his interests. $P_v = \{tw_1, tw_2, ..., tw_n\}$ is the total posts created by the user ***v***. To determine the relationship between a post and the topic of interest to the user, TF/IDF method and cosine of the angle between vectors of the words t_{wi} and P_v are used. Owing to the diversity of words employed, this method has low accuracy. For this purpose, different themes can be categorized in the texts using Latent Dirichlet Allocation. Each theme includes a set of words $M_{topic\#} = \{w_1, w_2, ..., w_L\}$, of which the probability of occurrence is specified in the relevant theme. To increase the accuracy of thematic similarity of candidates to previous posts of the user ***v*** ($P_v = \{tw_1, tw_2, ..., tw_n\}$), the angle between these two vectors is measured through the cosine, using Equation 4.

$$topsim(tw_i, P_v) = \frac{M_{topic_{tw_i}} . M_{topic_{P_v}}}{\| M_{topic_{tw_i}} \| . \| M_{topic_{P_v}} \|} \quad (4)$$

In this article, in addition to calculating the lexical similarity between the post tw_i and the posts used by the user, geographical similarity is taken into account. Normally, the influence of the posts addressing regional issues is higher than the Tweets of other regions. Therefore, in this article, maxmind data set was used including 4 million names of cities and regions along with additional information of the country, location, etc. to find the words related to cities and geographic regions. To find words related to cities from the post tw_i, all the words in the post but additional words are used (even the hashtags, the symbol # excluded). If Loc_{tw_i} be the set of cities plus country and the region used in the post and Loc_{u_i} be the set of cities, countries and regions used by the user u_i in all the posts created, Equation 5 shows the geographical relationship between the post tw_i and the user u_i.

$$Locsim(tw_i, u_i) = \frac{|(Loc_{tw_i} \bigcap_{u_i \in U \,\&\, tw_i \in P} Loc_{u_i})|}{|(Loc_{tw_i} \bigcup_{u_i \in U \,\&\, tw_i \in P} Loc_{u_i})|} \quad (5)$$

B- Similarity Between Users (User - User)

This criterion is very important for OCs. People with different specializations share their posts in the network. It is very beneficial to find users with common fields. For both users $u, v \in V$, the degree of similarity is equal to the degree of similarity between the posts already created. The essential thing about sharing information in OCs is to find people with the same level of information in addition to similar posts. For example, a user who has created more than 100 posts about smart phones applications is different from someone who has just had a few posts or reposts in the same field. For either user, the action vector can be defined (Equation (6)). This vector contains n keywords created or reposted by the user $v \in V$. Weighted cosine is used to determine the similarity between the vectors of users. In this respect, the coefficients i (number of keyword repeated by the user $v \in V$) is determined for n keyword till time t. Considering the coefficients i, the level of users' knowledge on a specific area is determined according to the number of posts made by them. The two users are examined and taken into account for determining the similarity given the repetition of the keywords in the posts.

$$vector_u^t = i_1^t keyword_1 + ... + i_n^t keyword_n \quad (6)$$

The degree of similarity between the users u and v is equal to the value of cos for vectors of these two users.

$$sim(u,v) = \cos(vector_u, vector_v, t) = \frac{vector_u^t \, vector_v^t}{||vector_u^t|| . ||vector_v^t||} \quad (7)$$

In equation (6), $||vector_v^t||$ or/and $||vector_u^t||$ respectively represent the value of action vector for the users v and u. If any of these values is zero, it means that the relevant user has had no action (not created nor reposted). In that case, the similarity between two users is not defined. Accordingly, in collecting data, only those users are taken into account who have at least created one post or reposted. Based on this assumption, there is no problem in calculating this similarity. It should be noted that the action vector of a user changes as the time changes. In this respect, the action vector of users at the time t is used in each determination of the similarity between two users.

C- Using Logistic Regression to Predict the Probability of Reposting

Logistic regression is used to estimate the probability of reposting. Typically, based on input features in logistic regression, a linear function is defined which, based on these features (similarity of candidate post, geographical similarity, and similarity between user and user), predicts the influence (being reposted) of a post using the sigmoid (Logistics) function (Equation 8).

$$y_i = h(w^T x_i) = \frac{1}{1 + \exp(-w^T x_i)} \quad (8)$$

In this equation, y_i is the prediction based on the x_i inputs. w^T is the vector of coefficients of each feature obtained from training data. Equation 8 shows the error function of logistic regression algorithm. In this equation, N is the number of reposts used in the training data set.

$$J(w^T) = \frac{1}{N} \sum_{i=1}^{N} \text{Cost}(h_w(x_i), y_i) \quad (9)$$

In Equation 9, $\text{Cost}(h_w(x_i), y_i)$ is the algorithmic cost function. Since binary classification algorithm is used, $J(w^T)$ is written as Equation 10. To obtain the weight of each feature, (w^T) should be determined based on the input data x_i and real data related to the users' repost in the data set y_i in such a way that $J(w^T)$ is minimized ($\min_{w^T} J(w^T)$).

$$J(w^T) = \frac{1}{N}[\sum_{i=1}^{N} y_i \log h_w(x_i) + (1-y_i)\log(1-h_w(x_i))] \quad (10)$$

To cope with over fitting in the equation above, regularization term is taken into account. The value of error function regarding this regularization term is given in the Equation (11). *m* is the number of features used for classification.

$$J(w^T) = \frac{1}{N} \times \sum_{i=1}^{N} y_i \log h_w(x_i) + (1-y_i)\log(1-h_w(x_i)) + \sum_{j=1}^{m} (w_j)^2 \quad (11)$$

The reason behind using logistic regression is its capability with respect to the issue mentioned above. In addition to classification, this method has also probable output. For example, $h_w(x_i) = 0.8$, that is, the probability of retweet is at 80% for the sample ($h_w(x_i) = p(y_i = 1 | x_i; w^T)$, which is very useful for the question in this article considering the need to estimate the probability of reposting.

3.3 Probability of Reposting Based on User Type

The calculated probability of repost of post p of the user u by the user v at timestamp t (put formally) $p(repost(v,u,p,t))$ is shown in the algorithm (1). In algorithm (1), t_u is the time a post is reposted by the user u. t is the run-time of the algorithm and γ_v is the average time interval between the posts of user v. α is the adjusted coefficient of γ_v. At the zero line of this algorithm by calculating the elapsed time since the repost of the posts by the user u, if the time is less than the

average time interval between the posts of the user v ($\alpha\gamma_v$), there is still the probability that the user v has not seen the post p. In this respect, the probability of reposting by the user v is equal to max($p^p_{u,v}, \varepsilon$). The line 2 evaluates a condition in which the time elapsed since the repost of the posts by user u is longer than the average time interval between the posts of the user v ($\alpha\gamma_v$). In this case, the user had most likely seen the post but was unwilling to repost it. Considering this fact, it is least probable that the user v repost the post p of user v in the timestamp t ($p(repost(v,u,p,t)).\varepsilon$ is equal to the very minimal value at 10^{-3}.

Algorithm (1) calculating the probable repost of post p of the user u by the user v at timestamp t	
Input: t_u is the time the post is shared by user *u*, the present time *t*, γ_v, $p^p_{u,v}$	
Output: *p*(*repost* (*v, u, p, t*)) probable repost of post p of the user *u* by the user *v* at timestamp *t*	
Definitions: ε is the minimal value, γ_v is the average time interval between the posts of the user *v*, $p^p_{u,v}$ is the probable repost of post *p* (from posts of u) by the user v, α is the coefficient	
00	If $t\text{-}t_u <= \alpha\gamma_v$
01	*p*(*repost* (*v, u, p, t*))=max($p^p_{u,v}, \varepsilon$)
02	if otherwise
03	*p*(*repost* (*v, u, p, t*))=ε
04	End

Accordingly, if the time elapsed since the creation of the post is longer than the time the user is inactive in the online community, it is least likely to be displayed and its value is equal to ε . The coefficient of α allows more opportunity for lurking users by adjusting the impact of γ_v. Normally, γ_v of lurking users is far longer than γ_v of other users. For ease of calculation, the coefficient α of γ_v is considered to be 1 and 2 for ordinary users and lurking users, respectively. In this respect, when a user is lurking, he/she has more time to read and repost the posts with higher probability.

4. Results

In this section, we will discuss in detail the experimental data and simulation framework as well as the results of this research.

4.1 Experimental Data & Simulation Framework

Twitter data were used to test the proposed method. With more than 200 million users, Twitter witnesses a million posts per day. This social networking site is a directed graph including users who freely follow other users. Each user is able to create a post called “Tweet”. These tweets contain photos, URL links, texts and so on. Each tweet contains a maximum of 140 words. Tweets are registered in the user page after they are created and can be viewed by the users who follow the former ones. Re-Tweets refers to the repost of a user’s Tweet by other users. Each user Retweets a Tweet according to his interest in it. Twitter as a huge data base is in the focus of many researchers in this field [7], [30], [31]. For collecting the data of Twitter Search API, all public tweets related to the field of IT technology with different keywords (8 keywords) were extracted from September 20 to December 24 of 2011. The data contains 94 thousand tweets. By making a Twitter API query for each user ID, its metadata including the followers and the followees were extracted. To map the social graph of users, only those users were taken into account who had at least Tweeted or Retweeted a post.

For testing, the simulation method similar to [7], [32], [33] was used. All the parameters and variables are based on actual data. On obtaining the social network graph and the set of tweets as input, the probability of reposting $P(repost(v,u,p,t))$ was calculated. Then, reposting is begun randomly and based on the model of post sharing. In order to carry out a simulation based on real data, an hour scale is considered for the time in the proposed method. This makes it possible to select k influential posts for each user per hour. First, the simulation is started from an augment node for λ followed by all nodes. This makes it possible to uniformly enter the posts of the data set. The rate at which posts are entered into the simulation environment acts as an input parameter. The other assessment input parameter is the number of iterations occurred for the simulation. In general, simulation is started by displaying the posts created by λ for each user at any timestamp t. Then, by calculation of $\sigma(v,t)$ using the proposed method, the k influential posts are determined for the user v. At each attempt, λ begins to uniformly create posts for users. Depending on the model of reposting, the users begin to randomly repost the posts (taking into account the probability of biased reposting), as was mentioned earlier.

4.2 Evaluation of Post Selection Method Using Simulation

To test the proposed method, 72 hours simulation was carried out. In each hour simulation for the first 10 hours, λ created 10 posts on average. In the first 10 hours, a total of 100 Twitter posts were examined. The social graph of the data collected contains 4.9 thousand users. A variety of implementations was done for various modes, to be evaluated as follows.

, to evaluate the increasing of information propagation in OC, 4 more methods, other than the proposed one, are also employed. 1) Latest post method: based on the time a post has created, the learners visit the latest k post. 2) Random k post selection method for a learner. 3) Post-post similarity method: The k posts with the highest similarity to the learner’s posts. 4) learner-Learner similarity method: k post selection method with the highest similarity of the posts authors with the learner.

As shown in Table (1), the results of reposting activities done according to different methods is given.

The Random refers to a method that randomly selects k posts from among candidate posts for display of the node v. Similarly, Recency (Rec) selects the recent k posts from candidate posts for display. These two basic methods in [8], [19] are used for comparison. As is clear from the graph, the social similarity method in the data set under study had the best performance compared to the similarity of users. It should be mentioned that the method of post-to-post similarity performed better than the method of individual's influence. As predicted, the two Random and Recency methods had the worst performance for k=5 (selection of five influential posts for each user).

Table 1 shows the results for the two modes k = 3 and k = 8. As specified in the table, Random method had no acceptable performance for k = 8. This may be due to low rate of post entry. This method cannot be efficient for high rates of posts. The results in k = 5 are roughly the same for k = 3.

Table 1. Repost action per thousand for two selection modes of 3 and 8 influential posts.

k=8	***k=3***	***Methods***
activity (1000=K)	***Activity (1000=K)***	
1.5	3	***Random***
3.6	2.3	***Recent***
6.8	2.9	***Post-Post Similarity***
6.9	2.7	***User-User Similarity***
8.2	4.1	***Proposed Method***

4.3 Repost Prediction Method Evaluation

The correlation of these features with repost behavior is studied in this paper. If the values of these features are significantly related to repost behavior, the relationship can be measured based on conventional learning. To this end, a method similar to Pearson's correlation method is employed. Since each feature has continuous values and the repost behavior is a binary variable (0 or 1), Pearson's method cannot be used. The point-biserial method is utilized for such problems. If the values of each feature of each post are a continuous variable (x), and repost behavior for the post is a binary variable (y), then the point-biserial correlation coefficient is based on formula (11), where M1 shows the mean feature value for posts, leading to repost behavior y=1. Similarly, M0 is the mean of features of posts, not leading to repost behavior y=0. Moreover, n1 shows the number of posts in the samples except for posts leading to repost behavior (y=1), and n0 denotes the number of posts in samples which do not result in repost behavior (y=0). In addition, n is the total number of samples examined, sn is the standard deviation for values of features of all of the studied post samples. This correlation coefficient varies between -1 and 1, where 1 shows maximum positive correlation between measured values and user behavior and -1 shows maximum negative correlation between measures values and user behavior. Zero (0) also shows independence of the features from user repost behavior.

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{11}$$

Table (2) presents results of the coefficient of correlation between features and reposts behavior. These results are reflective of a positive correlation between these features and repost behavior. The maximum correlation belongs to post-post similarity.

Table 2. Coefficient of correlation between proposed features and repost behavior

	Post-Post Similarity	User-User Similarity
r_{pb}	0.63	0.54

Two data sets from the Twitter social media were used to assess the proposed method. The data was selected based on the following selection criteria: 1) All users should have more than 30 items in their list of followers and followee; 2) Users creating or retweeting at least 20 tweets a week. Each of the selected data sets has a one-month history of general tweets and meets the mentioned criteria. These two data sets were selected twice in two weeks in 2010 and 2011. These two data sets were selected because due to the content-based nature of the proposed method, it was tried to conduct assessments when the topics were not the same. Finally, after preprocessing procedures, a total of 23038 active users were identified in the two data sets, which contained 1211347 tweets and 92307 retweets. Table (3) presents overall specifications of the data sets collected for assessment.

Table 3. Data sets specifications

Data Set 1		***Data Set 2***	
Tweet	Retweet	Tweet	Retweet
787376	62191	423971	30116

Three well-known measures, namely "precision", "recall", and "F measure", are used to assess the predictions [12], [13]. These measures are used for prediction problems of the binary class. Formulas 12 to 14 show these measures.

$$Precision = \frac{|\{\text{Predicted RT}\} \cap \{\text{True RT}\}|}{|\{\text{Predicted RT}\}|} \tag{12}$$

$$\text{Recall} = \frac{|\{\text{Predicted RT}\} \cap \{\text{True RT}\}|}{|\{\text{True RT}\}|} \tag{13}$$

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Using the conventional supervised classification method, the data sets were divided into the training and test groups. Table (4) presents specifications of the training and test groups of the assessment data sets.

Table 4. Specifications of test and training data sets

	Data Set 1		Data Set 2	
	Tweet	Retweet	Tweet	Retweet
Training	623340	49235	335643	23841
Test	164036	12956	88328	6275

For the purpose of binary classification the decision tree method, Naïve Bayes and proposed logistic regression method were used as the bases. Table (5) presents analysis results of the two data sets.

Table 5. Experiment results

	Data set 1			*Data Set 2*		
	Precision	***Recall***	***F measure***	***Precision***	***Recall***	***F measure***
Proposed Method	0.941	0.68	0.789	0.88	0.61	0.720
Naïve Bayes	0.92	0.65	0.763	0.867	0.51	0.702
Decision tree	0.79	0.54	0.746	0.832	0.46	0.593

5. Limitation & Conclusion

The most obvious limitation of this study was the absence of a specific standard in the implementation and design of OCs. Therefore, the suggested method is designed based on the content of posts to make it applicable in various kinds of OCs. Other factors such as external events (such as trends and news), context of the OC (such as health forums or technical forums), security concerns and OC design problems can be effective in users' repost behavior. This article seeks to solve the problem of selecting k influential posts for the user v from among the posts a user like u has created and the user v follows. These posts should be selected in such a way that the entire reposts increase in the network. On reviewing the literature in this field, the complexities of the issue were discussed. The existing methods have failed to be implemented online so far. Since this issue has been defined for knowledge management in online environments, these methods were not usable. In addition to online implementation, the proposed method is suitable for use in OCs. Certain features such as post-post similarity, user-user similarity, and geographic similarity are among major parameters taken into account in the selection of k influential posts. The proposed method focuses on the problem of online communities that is lurking users. An evaluation based on real data and different scenarios makes the proposed method more efficient compared to other methods.

References

[1] W. Guechtouli, J. Rouchier, and M. Orillard, "Structuring knowledge transfer from experts to newcomers," J. Knowl. Manag., vol. 17, no. 1, pp. 47–68, 2013.

[2] M. G. Wilson, J. N. Lavis, R. Travers, and S. B. Rourke, "Community-based knowledge transfer and exchange: Helping community-based organizations link research to action," Implement. Sci., vol. 5, no. 1, p. 33, 2010.

[3] C. K. K. Chan and Y. Y. Chan, "Students' views of collaboration and online participation in Knowledge Forum," Comput. Educ., vol. 57, no. 1, pp. 1445–1457, 2011.

[4] B. Nonnecke and J. Preece, "Lurker demographics: Counting the silent," Proc. SIGCHI Conf. …, vol. 2, no. 1, pp. 1–8, 2000.

[5] B. Nonnecke and J. Preece, "Why lurkers lurk," AMCIS 2001 Proc., pp. 1–10, 2001.

[6] P. G. Kilner and C. M. Hoadley, "Anonymity options and professional participation in an online community of practice," Proc. 2005 Conf. Comput. Support Collab. Learn., pp. 272–280, 2005.

[7] Y. Amichai-Hamburger, T. Gazit, J. Bar-Ilan, O. Perez, N. Aharony, J. Bronstein, and T. Sarah Dyne, "Psychological factors behind the lack of participation in online discussions," Comput. Human Behav., vol. 55, pp. 268–277, 2016.

[8] P. Sloep and L. Kester, "From lurker to active participant," in Learning Network Services For Professional Development, 2009, pp. 17–25.

[9] M. Takahashi, M. Fujimoto, and N. Yamasaki, "The active lurker: influence of an in-house online community on its outside environment," Group, pp. 1–10, 2003.

[10] P. Kollock and M. Smith, "Managing the Virtual Commons: Cooperation and Conflict in Computer Communities," Comput. Commun. Linguist. Soc. Cross-Cultural Perspect., pp. 109–128, 1996.

[11] A. Schneider, G. Von Krogh, and P. J??ger, "What's coming next? Epistemic curiosity and lurking behavior in online communities," Comput. Human Behav., vol. 29, no. 1, pp. 293–303, 2013.

[12] H. Zhang, Q. Zhao, H. Liu, K. Xiao, J. He, X. Du, and H. Chen, "Predicting retweet behavior in Weibo social network," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7651 LNCS, pp. 737–743.

[13] X. Tang, Q. Miao, Y. Quan, J. Tang, and K. Deng, "Predicting individual retweet behavior by user similarity: A multi-task learning approach," Knowledge-Based Syst., vol. 89, pp. 681–688, 2015.

[14] J. Leskovec, L. Adamic, and B. Huberman, "The Dynamics of Viral Marketing," ACM Trans. Web, vol. 1, no. 1, pp. 1–39, 2007.

[15] E. Sun, I. Rosenn, C. a Marlow, and T. M. Lento, "Gesundheit! Modeling Contagion through Facebook News Feed Mechanics of Facebook Page Diffusion," Proc. Third Int. ICWSM Conf., no. 2000, pp. 146–153, 2009.

[16] E. Bakshy, B. Karrer, and L. A. Adamic, "Social Influence and the Diffusion of User Created Content," in Electronic Commerce, 2009, pp. 325–334.

[17] P. Domingos and M. Richardson, "Mining the Network Value of Customers," in Proceedings of the Seventh {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, 2001, pp. 57–66.

[18] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 02, vol. 02, no. 3, p. 61, 2002.

[19] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03, 2003, p. 137.

[20] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," Proc. VLDB Endow., vol. 5, pp. 73–84, 2011.

[21] U. Feige, V. S. Mirrokni, and J. Vondrák, "Maximizing non-monotone submodular functions," in Proceedings -

Annual IEEE Symposium on Foundations of Computer Science, FOCS, 2007, pp. 461–471.
[22] S. Stieglitz and L. Dang-Xuan, "Emotions and Information Diffusion in Social Media - Sentiment of Microblogs and Sharing Behavior," J. Manag. Inf. Syst., vol. 29, no. 4, p. 217, 2013.
[23] S. Ye and S. F. Wu, "Measuring message propagation and social influence on Twitter.com," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6430 LNCS, pp. 216–231.
[24] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6913 LNAI, pp. 18–33.
[25] M. Magnani, D. Montesi, and L. Rossi, "Information Propagation Analysis in a Social Network Site," 2010 Int. Conf. Adv. Soc. Networks Anal. Min., pp. 296–300, 2010.
[26] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Effects of user similarity in social media," in Proceedings of the fifth ACM international conference on Web search and data mining WSDM 12, 2012, p. 703.
[27] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," Proc. third ACM Int. Conf. Web search data Min. - WSDM '10, p. 241, 2010.
[28] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," The Journal of Mathematical Sociology, vol. 15. pp. 193–206, 1990.
[29] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social Influence Analysis in Large-scale Networks," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 807–816.
[30] M. Nagarajan, H. Purohit, and A. Sheth, "A Qualitative Examination of Topical Tweet and Retweet Practices," Artif. Intell., pp. 295–298, 2010.
[31] C. G. Knight and L. K. Kaye, "'To tweet or not to tweet?' A comparison of academics' and students' usage of Twitter in academic contexts," Innov. Educ. Teach. Int., no. April 2015, pp. 1–11, 2014.
[32] C. Haeussler, "Information-sharing in academia and the industry: A comparative study," Res. Policy, vol. 40, no. 1, pp. 105–122, 2011.
[33] W. S. Hwang, S. W. Kim, D. H. Bae, and Y. J. Do, "Post ranking algorithms in blog environment," in Proceedings of the 2008 2nd International Conference on Future Generation Communication and Networking, FGCN 2008, 2008, vol. 2, pp. 64–67.

Omid Reza Bolouki Speily received the B.Sc. degree in Computer Engineering from Urmia University, the M.Sc. & Ph.D. degrees in Information Technology from the AmirKabir University of Technology. He worked as a researcher at the Iran Telecommunication Research Center (ITRC). Since 2009 he joined the Urmia University of Technology as a faculty member of Information Technology & Computer Engineering Department. His research interest includes the dynamics complex networks, graph theory, intelligent system, e-Services.

Good Index Choosing for Polarized Relay Channel

Hassan Tavakoli*
Department of Electrical Engineering, University of Guilan, Rasht, Iran
htavakoli@guilan.ac.ir
Saeid Pakravan
Department of Electrical Engineering, University of Guilan, Rasht, Iran
Saeidpak70@yahoo.com

Received: 08/Nov/2016 Revised: 02/Sep/2017 Accepted: 20/Sep/2017

Abstract

The Polar coding is a method which have been proposed by Arikan and it is one of the first codes that achieve the capacity for vast numerous channels. This paper discusses relay channel polarization in order to achieve the capacity and it has been shown that polarization of two relay channels can be given a more achievable rate region in the general form. This method is compatible with the original vision of polarization based on the combining, splitting and polarizing of channels and it has been shown that the complexity of encoding and decoding for these codes in mentioned method are $O(N\,log\,N)$, and also error probability for them is $O(2^{-(N)^{\beta}})$. Choose the best sub-channels in polarized relay channels for sending data is a big trouble in this structure. In this paper, we have been presented a new scheme for choosing a good index for sending the information bits in relay channels polarized in order to have the best performance by using sending information bits over FIF sets.

Keywords: Relay channel; Polar code; Channel polarization; Capacity; Relay channel polarization; Good index of relay channel.

1. Introduction

The relay channel is a communication channel which has a sender and receiver assistant in communication by utilizing of a relay node [1]. Specifying a memoryless relay channel is can be given by the probability distribution $W(Y_r, Y|X, X_r)$. In the defined prbability, , X_r are the symbol transmitted by the source and the symbol transmitted by the relay, respectively, are Y_r is the symbol received by the relay and finally Y is the symbol received by the destination. This defination has been illustrated generally in Fig. 1. In this model of relay channel, it has been assumed that the message M is uniformly distributed throughout the message set and the average probability of error is defined as

$$P_e^{(n)} = \Pr\{\hat{M} \neq M\} \tag{1}$$

The rate R is said to be achievable if there is a message with $(2^{NR}, N)$ codes property such that

$$\lim_{N\to\infty} P_e^{(n)} = 0 \quad . \tag{2}$$

In [2], it is well-known and it has been shown in [1], which the capacity of the relay channel in general form is still an open problem and this causes the inportance of relay channel study.

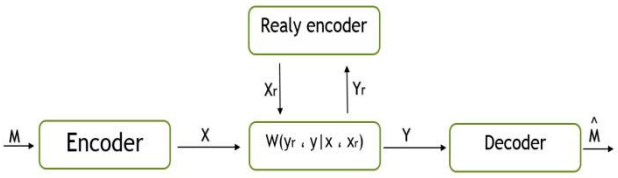


Fig. 1. A typical relay channel [1]

The cut-set bound is the outer bound of the capacity for the relay channel and it is established in [2] as follows:

$$C \le \max_{p(x,x_r)} \min \{I(X, X_r; Y_r), I(X; Y, Y_r | X_r)\}. \tag{3}$$

Decode-and-forward (DF) and compress-and-forward (CF) are the main coding scenarios for information transmitting in relay channels. In DF strategy, after recoverying the transmitted message from sender by relay, the relay forward it to the destination and this information helps the receiver to complete the best observation of the main link. Lower bound of DF is given by [2]:

$$C \ge R_{DF} := \max_{p(x,x_r)} \min \{I(X, X_r; Y_r), I(X; Y_r | X_r)\}, \tag{4}$$

Recently, polar codes, introduced by Arikan, give a way that can be called channel polarization technique and this scheme have been extended to various multi-terminal scenarios, such as the Multiple-Access Channels [5-7], Broadcast Channels [8]-[9] as well. These codes are one of the first codes that can achieve the capacity for binary input symmetric channels [3]-[4].

* Corresponding Author

In this paper, firstly, the well-known channel polarization phenomenon has been presented for relay chnnel and has been shown that the polarization of cut-set bound in relay channels how can be used effectively. The proposed schemes have the same standard properties of a typical polar codes with respect to encoding and decoding with the complexity $O(N.logN)$. The scaling of the block error probability is an exponential function of the block length, which decay like $(2^{-(N)^{\beta}})$, where $O \leq \beta \leq 0.5$. The choosing the best sub-channels and good index in polarized relay channels for sending the data is a big trouble problem. In our structure, it has been solved by using sending the information over an FIF set in indices at Section V.

This paper is organized as follows: In Section II, the backbone material of a typical polar code and relevant previous works on relay channels have been reviewed. It has been shown also that the DF and CF strategy in general three terminal relay channel with using polar codes are achievable in Section III. In Section IV, polarization of relay channel for cut-set bound has been shown. In Section V, we introduce a new scenario of choosing good indices for relay channels polarization and finally. Last but not the least, at Section VI, we conclude the paper.

2. polar codes and relay channels

In this section, we supply a brief overview of the main work of Arikan [3] for single link polar codes technique and channel polarization method we present a brief part of previous works about the relay channels.

A. Polar codes

Constructing a typical polar codes is based upon a phenomenon that is well-known as channel polarization method [4]. The basic channel polarization method is given by a matrix, which is:

$$G_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{5}$$

The kronecker power of G for any $n \succ 1$ can be defined according to the iterative matrix formula such as

$$G_2^{\otimes n} = \begin{bmatrix} G_2^{\otimes(n-1)} & 0 \\ G_2^{\otimes(n-1)} & G_2^{\otimes(n-1)} \end{bmatrix} \tag{6}$$

with initial point as (5). Following [3], for a typical DMC, the channel splitting can be defined as mapping of

$$(W, W) \to (W^-, W^+). \tag{7}$$

The synthesized channels

$$W^- : F_2 \to Y^2 \tag{8}$$

and

$$W^+ : F_2 \to F_2 \times Y^2 \tag{9}$$

are given as follows:

$$W^-(y_1^2 \,|\, u_1) = \sum_{u_2 \in \{0,1\}} \frac{1}{2} W(y_1 \,|\, u_1 \oplus u_2) W(y_2 \,|\, u_2), \tag{10}$$

$$W^+(y_1^2, u_1 \,|\, u_2) = \frac{1}{2} W(y_1 \,|\, u_1 \oplus u_2) W(y_2 \,|\, u_2). \tag{11}$$

W^- and W^+ are bad and good channel, respectively, in comparing to the original channel by using the channel splitting method. Arikan uses the Bhattacharyya parameter for a typical channel W, which is denoted by $Z(W)$, in order to measure how a good channel and bad channel can be classified. This parameter in general form is defined by:

$$Z(W) = \sum_{y \in Y} \sqrt{W(y \,|\, 0) W(y \,|\, 1)}. \tag{12}$$

Channels with Bhattacharyya parameter close to "0" are almost noise-free channel and channels with $Z(W)$ close to "1" are almost full-noisy channels [4]. The sub-set of noise-free channels $I_N(W)$ is defined according to below for any $O \leq \beta \leq 0.5$:

$$I_N(W) := \{i \in [N] : Z(W_{2N}^{(i)}) \leq \frac{2^{-(N)^{\beta}}}{N}\}. \tag{13}$$

Where [*N*] is denoted the set of channels, which has less or equal to *N*. The channel polarization technique certify that the fraction of good binary input channel, approaches the symmetric capacity *I(W)* when *N*approaches to infinity [3]-[4]. *I(W)* is the quantity of mutual information and it can be defined as the channel capacity of *W*. By using the polarization theorem, one can find out that polar codes achieves the capacity [3]. Let set *A* is characterized as

$$A = \{i \in [N] : Z(W_N^i) \in [0, \delta]\}, \tag{14}$$

where $\delta \succ 0$. For any $O \leq \beta \leq 0.5$, the error probability of polar codes is determined by block error-probability when the Successive Cancellation (SC) decoding used:

$$P_e = \sum_{i \in A} Z(W_N^{(i)}) = o(2^{-2^{n\beta}}). \tag{15}$$

B. Previous works on relay channels with respect to polar coding:

There are two ideas in using of polar codes. The main idea, which was shown by Arikan, is that the polar codes can achieve the capacity for a large case of channels and it is equal to say that the rate of such scheme approaches the capacity of the channel. Also, the main idea is still capacity achieving polar codes in most papers about the relay channel. In [10], a practical method for acheiving the capacity of symmetric physically degraded relay

channels with binary input has been shown effectively. By concluding [11], one can find out about the achievability of DF bound by means of polar codes in a stochastically channel, which is degraded with binary symmetric relaying and orthogonal receivers. Also in [11], it has been shown that polar codes can be applied to CF relaying. In [12-13]. it has been discussed that by utilizing the polar codes, one can achieve the capacity of symmetric degraded relay channel. Also the problem of achievability for two other lower bounds by utilizing polar coding techniques has been shown in [14]. The topic of increasing the capacity for $N \to \infty$ has been studied in [5-7] and it has been shown that this the capacity of one of the polarized channels increase while the other decrease. In [5-6], a method for the polarization of the MAC has been illustrated. It has been shown that the polarization of a general Multiple Access Channel with a point-to-point channel can be done and one can achieve more rate region [7].

In second approach, one can design a channel polarization scheme when the number of channels increases, N and show that the capacity region increases too. Changing of different bound of relay channel, especially cut-set bound is a fantastic vision at this approach. In this paper, we use the first and the second methodology together in order to choose good index for the relay channel polarization method.

3. DF and cf relaying using polar codes

In [10]-[13], for degraded relay channel, polar coding schemes have been proposed on the following Markov chain as:

$$X \to (X_R, Y_R) \to Y \tag{16}$$

Also, a polar coding scheme can be applied for compress-forward and it has been proposed in relay channels with orthogonal receiver components as well in [11].

Theorem1: based on the first viewpoint, (17), (18) give the following relation:

$$\mathrm{I}(\mathrm{W}^-) + \mathrm{I}(\mathrm{W}^+) = 2\mathrm{I}(\mathrm{W}), \tag{17}$$

and

$$\mathrm{I}(\mathrm{W}^-) \le \mathrm{I}(\mathrm{W}) \le \mathrm{I}(\mathrm{W}^+). \tag{18}$$

Where W^- and W^+ are bad and good channels, respectively.

Proof: The proof is the same as [3], [4].

The mapping of

$$(\mathrm{W}, \mathrm{W}) \to (\mathrm{W}^-, \mathrm{W}^+) \tag{19}$$

has been called polarization (one level polarization). The same mapping can be applied to W^- and W^+ to get $\mathrm{W}^{--}, \mathrm{W}^{-+}, \mathrm{W}^{+-}$ and W^{++} (which is second level polarization). For any arbitrary number of levels, the same process can be continued in order to polarize W. The channel polarization theorem states that the binary input channel can be polarized as the code length N goes to infinity, it means that they can be set in two different sets, one set become noise-free and the other become very noisy. In this viewpoint instead of

$$\mathrm{R}(\mathrm{W}) \le \mathrm{I}(\mathrm{W}), \tag{20}$$

by using (18) more rate can achieve, so:

$$I(W) \le R(W^+) \le I(W^+) \tag{21}$$

We call this vision as channel polarization in order to achieve more capacity.

In second viewpoint, since

$$\mathrm{I}(\mathrm{W}^+) \ge \mathrm{I}(\mathrm{W}), \tag{22}$$

we can find out a way, which

$$\mathrm{R}(\mathrm{W}^+) \ge \mathrm{R}(\mathrm{W}). \tag{23}$$

In polar coding, a channel, which can send bits without noise, uses information bit (I), and a channel, which can send bits noisy, uses frozen bit (F) or redundancy. Indeed, the polarization idea has been used to propose polar codes and a recursive process leads to efficient coding structures (encoding and decoding structures).

4. polarization for relay channel

CF strategy for relay channels based on orthogonal receiver components, $\mathrm{Y} = (\mathrm{Y}', \mathrm{Y}'')$, can be used and one can applied the well-known polar coding scheme as:

$$W(y_r, y|x, x_r) = W(y', y_r|x)W(y''|x_r) \tag{24}$$

[11]. This viewpoint can achieve the symmetric CF rate. The main results for this section is given in the following theorem.

***Theorem 2*. (**Symmetric DF and CF relaying using polar code). For any transmission rate

$$R < R(DF) \tag{25}$$

and any fixed rate

$$R < R(CF), \tag{26}$$

one can find a polar codes with block error probability

$$P_e^{(n)} = Pr\{\hat{M} \neq M\} \tag{27}$$

under SC decoding. This block error probability is bounded as

$$P_e \le O\left(2^{-(N)^{\beta}}\right), \tag{28}$$

where $0 < \beta < \frac{1}{2}$ and the relay channel is stochastically degraded.

***Proof*.** This Proof is like the proof of Theorem 1 in [10] and Theorem 2 in [13].

Since the main results have been surveyed, only the structure of polar codes, can be achieved for N relay channel and the complexity of encoding and decoding is $O\ (N.\log N)$ [10], [13].

The notion of channel polarization has been extended to relay channels, wherein a technique is described to polarize a given binary-input relay channel same as in [3]-[5]-[6]. We prove polarization of cut-set bound. It has been shown after polarization for two relay channels, the capacity of one relay increases while the other decreases and the capacity region of the relay changes. Two independent uses of the channel W results relay channel W^2 according to the polarization of Fig. 1, we have:

$$X \to (X_1, X_2), \tag{29}$$

$$X_r \to (X_{r1}, X_{r2}) \tag{30}$$

and

$$Y \to (Y_1, Y_2). \tag{31}$$

The cut-set bound in a channel W^2, is described by two following quantities:

$$I(X_1X_2, X_{r1}X_{r2}; Y_1Y_2) = 2I_1(W) \tag{32}$$

and

$$I(X_1X_2; Y_{r1}Y_{r2}; Y_1Y_2|X_{r1}X_{r2}) = 2I_2(W) \tag{33}$$

Also the cut-set bound is as follows:

$$R < \min\{I(X_1^2, X_{r1}^2; Y_1^2), I(X_1^2; Y_1^2, Y_{r1}^2|X_{r1}^2)\}, \tag{34}$$

and we have;

$$X_1^2 = U_1^2\, G_2 \tag{35}$$

and

$$X_{r1}^2 = V_1^2\, G_2. \tag{36}$$

Now we get:

$$2I_1(W) = I(U_1U_2, V_1V_2; Y_1Y_2) = I(U_1V_1; Y_1Y_2) + I(U_2V_2; Y_1Y_2, U_1V_1) = I_1(W^-) + I_1(W^+) \tag{37}$$

and we also have:

$$2I_2(W) = I(U_1U_2; Y_{r1}Y_{r2}, Y_1Y_2|V_1V_2) = I(U_1; Y_{r1}Y_{r2}, Y_1Y_2|V_1V_2) + I(U_2; Y_{r1}Y_{r2}, Y_1Y_2, U_1|V_1V_2) = I_2(W^-) + I_2(W^+) \tag{38}$$

In this way, the polarized bad relay channel is:

$$U_1 \times V_1 \to Y_1Y_2Y_{r1}Y_{r2} \tag{39}$$

and the capacity indicates with both quantities: $I_1(W^-)$ and $I_2(W^-)$. At the other hand, the polarized good relay channel is:

$$\mathrm{U_2} \times \mathrm{V_2} \to \mathrm{Y_1Y_2Y_{r1}Y_{r2}U_1|V_1V_2} \tag{40}$$

and the capacity indicates with both quantities: $I_1(W^+)$ and $I_2(W^+)$. Let 's define the channel

$$W: X \times X_r \to Y \times Y_r \tag{41}$$

as a relay channel with $\{0,1\}$ input alphabet. Now, we also define two other relay channels as

$$W^-: X \times X_r \to Y^2 \times {X_r}^2 \tag{42}$$

and

$$W^+: X \times X_r \to Y^2 \times {Y_r}^2 \times X \times X_r, \tag{43}$$

in which we have:

$$W^-(y_1^2, y_{r1}^2|u_1, v_1) = \sum_{u_2 \in x, v_2 \in x_r} \frac{1}{4} W(y_1, y_{r1}|u_1 \oplus u_2, v_1 \oplus v_2) W(y_2, y_{r2}|u_2, v_2) \tag{44}$$

and

$$W^+(y_1^2, y_{r1}^2, u_1, v_1|u_2, v_2) = \frac{1}{4} W(y_1, y_{r1}|u_1 \oplus u_2, v_1 \oplus v_2) W(y_2, y_{r2}|u_2, v_2); \tag{45}$$

Where W^- and W^+ correspond to bad and good relay channels, respectively. For capacity bound, we get:

$$I_i(W^-) \le I_i(W) \le I_i(W^+)\ ;\ \mathrm{i}=1,2 \tag{46}$$

Correspondingly, for R, we get:

$$R(W^-) < R(W) < R(W^+) \tag{47}$$

***Example 1*:** Assume the polarization of Fig. 1 and consider all links are BEC with erasure probability $\varepsilon_1 = 0.5$ except the source to relay which is a BEC channel with parameter erasure probability ε_2. Now, we investigate three mentioned scenarios, through polarization of two relay channels. For specific case, if the link between source to relay is polarized but the link between the relay to destination is not polarized, then, we have:

$$C_1^+ = 0.75 + \min\{0.75, 1 - \varepsilon_2\}$$

and

$$C_1^- = 1.25;$$

While, if the link between source to relay is not polarized but the link between the relay to destination is polarized, then, we have:

$$C_2^+ = 0.5 + min\{0.5, 1 - 0.5\varepsilon_2\}$$

and

$$C_2^- = 1.$$

As C_K^+ are representative of the most capacity (good channel). Now we examine the appropriateness of these two bounds. If $O \le \varepsilon_2 \le 0.75$, then $C_1^+ \le C_2^+$; and if $O.75 \le \varepsilon_2 \le 1$, then $C_1^+ \le C_2^+$. When $\varepsilon_2 = 0.75$, then all of the C_K^+ s will be equal to each other. It is worth noting that C_0^- and C_0^+ are, for the link between source to relay is polarized and the link between the relay to destination is polarized too, then, we have:

$$C_0^+ = 0.75 + min\{0.75, 1 - 0.5\varepsilon_2\}$$

and

$$C_0^- = 0.75 + min\{0.75, 0.5 + 0.5\varepsilon_2\}$$

Which are always better than two other cases.

Remark1: It is worth-mentioning that by using the polarization we can find more capacity in comparison to the non-polarized case. For example in a point-to-point channel by using channel polarization, the added capacity is:

$$\mathrm{I(X_2; X_1Y_1Y_2) - I(X_2; Y_2)} > 0$$

where X_i is the i-th input and Y_j is the j-th output. In this way, for the proposed channel polarization method of relay channel, the same thing happens. In the other words, According to the capacity region bound of [15], the capacity of the proposed relay channel in Example1 is:

$$\mathrm{C} = \min\left\{1 - \frac{\varepsilon_2}{2}, 0.75\right\}$$

which leads $\mathrm{C} \leq 0.75$. In fair comparison to the Example 1, the lowest capacity of C_1^+ is 0.75, which shows the extension of the capacity by using polar codes.

Now suppose that W is a binary input relay channel. Let $\{B_n\}_{n\geq 1}$ be an *i.i.d.* uniform random variable valued in {-,+}, with $Pr(B_1 = -) = Pr(B_1 = +) = \frac{1}{2}$ and let us define a relay channel with valued random process $\{\mathrm{W_n}: \mathrm{n} \geq 0\}$ via

$$\mathrm{W_0} := \mathrm{W}, \quad \mathrm{W_n} := \mathrm{W}_{\mathrm{n-1}}^{\mathrm{B_n}}, \ \mathrm{n} \geq 1 \tag{47}$$

Further, we define random processes $\{I_{1\,n}: n \geq 0\}$ and $\{I_{2\,n}: n \geq 0\}$ such that:

$$\mathrm{I_{1\,n}} := \mathrm{I_1(W_n)}, \ \mathrm{I_{2\,n}} := \mathrm{I_2(W_n)} \tag{48}$$

Lemma 1. The random processes $\{\mathrm{I_1(W_n)}: \mathrm{n} \geq 0\}$ and $\{\mathrm{I_2(W_n)}: \mathrm{n} \geq 0\}$ are bounded martingale.

Proof: Since W_n is a binary input relay channel, $I_1(W_n)$ and $I_2(W_n)$ take values in [0,1]; hence, the mentioned processes are bounded. The martingale property is claimed from (47), (48). The process $(I_1(W_n), I_2(W_n))$ converges almost surely and the limit is

$$(\mathrm{I_{1\infty}}, \mathrm{I_{2\infty}}) = \lim_{\mathrm{n}\to\infty}\left(\mathrm{I_1(W_n)}, \mathrm{I_2(W_n)}\right) \tag{49}$$

Now, the following theorem is given for calculating the Bhattacharya parameter, which is used in relay channels polarization.

Theorem 3. For each given relay channel with $I_1(W) = I_{MAC}$ and $\mathrm{I_2(W) = I_{BC}}$ and for the *BC phase*, we have:

$$\mathrm{W_{BC}} \to \mathrm{W_{BC}^-}, \mathrm{W_{BC}^+}: \left\{\mathrm{Z^-(W)} \leq 2\mathrm{Z_{BC}(W)}, \mathrm{Z^+(W)} \leq \mathrm{Z_{BC}}^2\mathrm{(W)}\right\}, \tag{50}$$

and for *MAC phase* of the relay channel, we have:

$$\mathrm{W_{MAC}} \to \mathrm{W_{MAC}^-}, \mathrm{W_{MAC}^+}: \left\{\mathrm{Z^-(W)} \leq 2\mathrm{Z_{MAC}(W)}, \mathrm{Z^+(W)} \leq \mathrm{Z_{MAC}}^2\mathrm{(W)}\right\} \tag{51}$$

Proof: We use the fact that for any binary input discrete memory-less channel W, we have:

$$\mathrm{I(W) + Z(W)} \geq 1, \tag{52}$$

and

$$I(W)^2 + Z(W)^2 \leq 1 \tag{53}$$

[4]. In polarizing mode, using [3] we get:

$$\mathrm{W^-: Z^-(W)} \leq 2\mathrm{Z(W)} - \mathrm{Z(W)}^2, \tag{54}$$

$$\mathrm{W^+: Z^+(W) = Z(W)}^2; \tag{55}$$

and using [5, 7, 15] we can write:

$$Z^-(W) \leq 2Z(W) \tag{56}$$

and

$$Z^+(W) \leq Z(W)^2. \tag{57}$$

Since the relay channel is combined of two *MAC* and *BC* channels, and can be considered as a point-to-point channel. Also, regarding to [16-17], we can write below relations for *MAC* and *BC phase* of relay channel, respectively. For *BC phase*, we have:

$$\mathrm{W_{BC}} = \{\mathrm{Z_1^-(W)} \leq 2\mathrm{Z_{BC}(W)} \text{ for } \mathrm{W_{BC}^-} \text{ and } \mathrm{Z_1^+(W)} \leq \mathrm{Z_{BC}(W)}^2 \text{ for } \mathrm{W_{BC}^+}\}; \tag{58}$$

also, for *MAC phase*, we have:

$$\mathrm{W_{MAC}} = \{\mathrm{Z_2^-(W)} \leq 2\mathrm{Z_{MAC}(W)} \text{for } \mathrm{W_{MAC}^-} \text{ and } \mathrm{Z_2^+(W)} \leq \mathrm{Z_{MAC}(W)}^2 \text{ for } \mathrm{W_{MAC}^+}\} \tag{59}$$

So, in general, we have:

$$\mathrm{Z(W^-)} \leq \max\{\mathrm{Z_1(W^-), Z_2(W^-)}\} \leq \max\{2\mathrm{Z_{BC}(W)}, 2\mathrm{Z_{MAC}(W)}\}, \tag{60}$$

$$\mathrm{Z(W^+)} \leq \max\{\mathrm{Z_1(W^+), Z_2(W^+)}\} \leq \max\{\mathrm{Z_{BC}(W)}^2, \mathrm{Z_{MAC}(W)}^2\} \tag{61}$$

One can conclude that

$$I(W^-{}_{mac}) + I(W^+{}_{mac}) \leq 2I(W_{mac}) \tag{62}$$

and

$$I(W^-{}_{bc}) + I(W^+{}_{bc}) \leq 2I(W_{bc}) \tag{63}$$

for any relay channel that links become polarized, and

$$R \leq \min\{I(W_{bc}), I(W_{mac})\} = 1 - \max\{Z(W_{bc}), Z(W_{mac})\} \tag{64}$$

Although, one can find the capacity by solving an optimization problem like [18], but the polarization technique can increase the capacity as well like [7].

We consider a relay channel with orthogonal receiver component. It consists of three nodes; a source node (S), a relay node (R), and a destination node (D). This system has three directed transmission links: the channel between source and receiver is displayed by W_{SD}, the channel between source and relay is displayed by W_{SR} and the channel between relay and destination is displayed by W_{RD}. We consider all the channel links have been polarized with generator matrix G_n. The relay is stochastically degraded in our system. Therefore, W_{SD} is stochastically degraded with respect to W_{SR}. Hence, if A_{SD} and A_{SR} are good channel sets for polar codes of W_{SD} and W_{SR} channels, respectively [16-17].

Lemma 2. For any two discrete W_{SD} and W_{SR} memory-less channels, if W_{SD} is degraded regard to W_{SR}, then $A_{SD} \subseteq A_{SR}$.

Proof. The proof of this lemma is like the Theorem 1 of [10].

5. Good Index Chosing in Relay Channel

We will always assume that these indices are labeled from 1 to *N* and that the processing order of the successive decoder is the one implied by this labeling. In this section, we describe our schema for choose a good index to the place set information from between the links which are polarized in the relay channel. First, we represent a polar block of length *N* by a row vector as in Fig. 2-a, such that any link has block-length $= 2^n$. For *N*=4, it has been shown that which one of the links are good after polarization. Here we offer a plan for a square screen $N \times N$; This is the similar structure of the Fig. 2-c for showing indices between the polarized link. In the previous sections, we used the word index to refer to one of the synthetic channels which are created by the polarization process. Represents the index in the intersection points of a link in Fig. 2-c that half of them are good for place information bits and remainder are bad, which are proper for frozen bits.

Now consider the relay channel shown Fig. 1. As can be observed, links between SR, SD and RD are determined through W_{SR}, W_{SD}, W_{RD} channels, respectively. For link SD, suppose W_{SD} be a DMC with a binary input *X* and output *Y*. Fix a distribution P_x for the random variable *X* Let $U^{1:N} = X^{1:N}G_N$, where $X^{1:N}$ is a vector of *n* i.i.d. components are drawn according to P_x. Consider the sets $\mathcal{H}_x$ and $\mathcal{L}_x$ defined as:

$$\mathcal{H}_x = \{i \in [N] : Z(U^i \mid U^{1:i-1}) \geq 1 - 2^{-(N)^\beta}\}, \tag{65}$$

$$\mathcal{L}_x = \{i \in [N] : Z(U^i \mid U^{1:i-1}) \leq 2^{-(N)^\beta}\}. \tag{66}$$

For $i \epsilon \mathcal{H}_x$, the bit U^i is approximately uniformly distributed and independent of $U^{1:i-1}$. In addition, $\mathcal{H}_x$ and $\mathcal{L}_x$ is consisted such that:

$$\lim_{N\to\infty} \frac{|\mathcal{H}_x|}{N} = H(X), \tag{67}$$

$$\lim_{N\to\infty} \frac{|\mathcal{L}_x|}{N} = 1 - H(X). \tag{68}$$

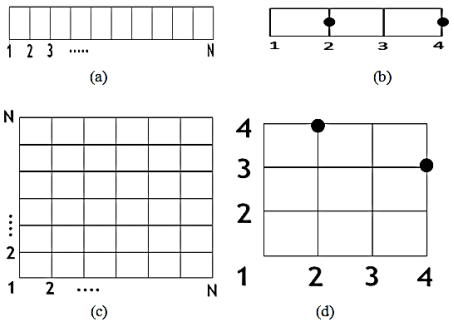


Fig. 2. a: A polar block of length *N* as a row vector. b: Good channels for polarized channel when *N*=4. c: A polar block of length $N \times N$ for each link in relay channel. d: A example for shown choose index in a channel with *N*=4.

Now, assume that the channel output $Y^{1:n}$ is given, and interpret this as side information on $X^{1:n}$. Consider the sets $\mathcal{H}_{X|Y}$ and $\mathcal{L}_{X|Y}$ as below:

$$\mathcal{H}_{X|Y} = \{i \in [N] : Z(U^i \mid U^{1:i-1}, Y^{1:N}) \geq 1 - 2^{-(N)^\beta}\}, \tag{69}$$

$$\mathcal{L}_{X|Y} = \{i \in [N] : Z(U^i \mid U^{1:i-1}, Y^{1:N}) \leq 2^{-(N)^\beta}\}. \tag{70}$$

For $i \epsilon \mathcal{H}_{X|Y}$, U^i is an approximately uniformly distributed and independent of $(U^{1:i-1}, Y^{1:n})$, and for $i \epsilon \mathcal{H}_{X|Y}$, U^i becomes approximately a deterministic function of $(U^{1:i-1}, Y^{1:n})$. Furthermore:

$$\lim_{N\to\infty} \frac{|\mathcal{H}_{X|Y}|}{N} = H(X|Y), \tag{71}$$

$$\lim_{N\to\infty} \frac{|\mathcal{L}_{X|Y}|}{N} = 1 - H(X|Y). \tag{72}$$

To create a polar code for the channel W_{SD}, we proceed now as follows: we place the information in the position indexed by

$$I_{SD} = \mathcal{H}_X \cap \mathcal{L}_{X|Y}. \tag{73}$$

Certainly, if $i \in I_{SD}$, then, U^i is approximately uniformly distributed given $U^{1:i-1}$, since $i \epsilon \mathcal{H}_x$. This implies that, U^i is suitable to contain information. Additionally, U^i given $U^{1:i-1}$ and $Y^{1:n}$, since $i \epsilon \mathcal{L}_{X|Y}$. Using (65)-(68) and the fact that the number of indices in $[N]$ which are neither in $\mathcal{H}_x$ nor in $\mathcal{L}_x$ is *O(N)*, it follows that:

$$\lim_{N\to\infty} \frac{|I_{SD}|}{N} = \lim_{N\to\infty} \frac{|\mathcal{L}_{X|Y} \setminus \mathcal{L}_X|}{N} = \lim_{N\to\infty} \frac{|\mathcal{L}_{X|Y}|}{N} - \lim_{N\to\infty} \frac{|\mathcal{L}_X|}{N} = H(X) - H(X|Y) = I(X;Y) = I_{SD}(W). \tag{74}$$

The remaining positions are frozen. More specifically, they are divided into two subsets, namely and $F_{d-SD} = \mathcal{H}_x^c$, that the frozen indices F_{r-SD} filled with binary bits $F_{r-SD} = \mathcal{H}_x \cap \mathcal{L}_{x|y}^c$, which are chosen uniformly at random and the frozen indices F_{d-SD}, which are chosen according to a deterministic rule. Similarly, to construct a polar code for the channel W_{SR}, we place the information on the positions indexed by:

$$I_{SR} = \mathcal{H}_{X|X_r} \cap \mathcal{L}_{X|X_r,Y_r} \tag{75}$$

and the remaining positions are related to frozen bits. They are divided into two subsets, namely

$$F_{r-SR} = \mathcal{H}_{X|X_r} \cap \mathcal{L}_{X|X_r,Y_r}^{\ c} \tag{76}$$

and

$$F_{d-SR} = \mathcal{H}_{X|X_r}^c, \tag{77}$$

which the frozen indices F_{r-SR} filled with binary bits selected uniformly at random and the frozen indices F_{d-SR} selected based on a deterministic principle. Since:

$$\mathcal{H}_{X|X_r} = \{j \in [N] : Z(U^j \mid U^{1:i-1}, X_r^{1:n}) \geq 1 - 2^{-(N)^\beta}\}, \tag{78}$$

$$\mathcal{L}_{X|X_r} = \{j \in [N] : Z(U^j \mid U^{1:i-1}, X_r^{1:n}) \leq 2^{-(N)^\beta}\}, \tag{79}$$

$$\mathcal{H}_{X|X_r,Y_r} = \{j \in [N] : Z(U^j \,|\, U^{1:i-1}, X_r^{1:n}, Y_r^{1:n}) \geq 1 - 2^{-(N)^\beta}\} \tag{80}$$

$$\mathcal{L}_{X|X_r,Y_r} = \{j \in [N] : Z(U^j \,|\, U^{1:i-1}, X_r^{1:n}, Y_r^{1:n}) \leq 2^{-(N)^\beta}\}. \tag{81}$$

So we have:

$$\lim_{N\to\infty}\frac{|I_{SR}|}{N} = \lim_{N\to\infty}\frac{|\mathcal{L}_{X|X_r,Y_r} \setminus \mathcal{L}_{X|X_r}|}{N} = \lim_{N\to\infty}\frac{|\mathcal{L}_{X|X_r,Y_r}|}{N} - \lim_{N\to\infty}\frac{|\mathcal{L}_{X|X_r}|}{N} = H(X|X_r) - H(X|X_r,Y_r) = I(X;Y_r|X_r) = I_{SR}(W). \tag{82}$$

Finally, for link RD, since

$$V^{1:n} = X_r^{1:n} G_N\,, \tag{83}$$

we place the information on the positions indexed by:

$$I_{RD} = \mathcal{H}_{X_r} \cap \mathcal{L}_{X_r|Y} \tag{84}$$

and the remaining positions are frozen. Based on the previous section, frozen bits related to this links are divided into two subsets F_{r-RD} and F_{d-RD}, which the frozen indices

$$F_{r-RD} = \mathcal{H}_{X_r} \cap \mathcal{L}_{X_r|Y} \tag{85}$$

filled with binary bits chosen uniformly at random and frozen indices

$$F_{d-RD} = \mathcal{H}_{X_r}^c \tag{86}$$

chosen according to a deterministic rule and since:

$$\mathcal{H}_{X_r} = \{k \in [N] : Z(V^k \,|\, V^{1:k-1}) \geq 1 - 2^{-(N)^\beta}\}, \tag{87}$$

$$\mathcal{L}_{X_r} = \{k \in [N] : Z(V^k \,|\, V^{1:k-1}) \leq 2^{-(N)^\beta}\}, \tag{88}$$

$$\mathcal{H}_{X_r|Y} = \{k \in [N] : Z(V^k \,|\, V^{1:k-1}, Y^{1:n}) \geq 1 - 2^{-(N)^\beta}\} \tag{89}$$

$$\mathcal{L}_{X_r|Y} = \{k \in [N] : Z(V^k \,|\, V^{1:k-1}, Y^{1:n}) \leq 2^{-(N)^\beta}\}. \tag{90}$$

So, we have:

$$\lim_{N\to\infty}\frac{|I_{RD}|}{N} = \lim_{N\to\infty}\frac{|\mathcal{L}_{X_r|Y} \setminus \mathcal{L}_{X_r}|}{N} = \lim_{N\to\infty}\frac{|\mathcal{L}_{X_r|Y}|}{N} - \lim_{N\to\infty}\frac{|\mathcal{L}_{X_r}|}{N} = H(X_r) - H(X_r|Y) = I(X_r;Y) = I_{RD}(W). \tag{91}$$

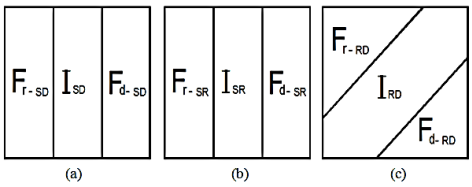


Fig. 3. Graphical representation of the sets associated to the channel coding problem (a. FIF plane for links SD b. FIF plane for links SR c.FIF plane for links RD).

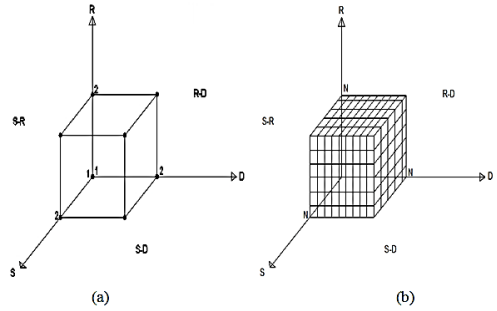


Fig. 4. Graphical representation of set index of relay channel (a. Set indices of relay channel with *N*=2 and b. Set indices for relay channel with N).

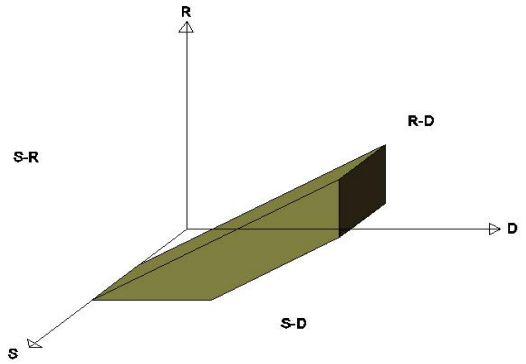


Fig. 5. Graphical representation of good channel in planes FIF

For Fig. 4-b, when $N \longrightarrow \infty$ intersection point become closer and closer to each other and in other words it can be displayed segmentation information as fig.5. When channels are polarize, synthesized channels can be classified into two categories, defining two index sets: the set information bits *I(w,g)* of indices corresponding to good channels and the set frozen bits *F(W)* of indices that belong to bad channels, and also set *F(W)* is consist of two part $F_d(w)$ and $F_r(w)$. For relay channel with *N=2*, it is shown that the set index according to Fig. 4-a., that each point in planes S, R, and D explain the links used in the relay channel. Similarly, Fig. 4-b. depicts relay channel with *N*, that, for example point *(i,j,k)*=(3,4,4) indicate in S-D, S-R, and R-D links 3, 4, and 4 are good respectively. For relay channel that $N \longrightarrow \infty$, set information of good channel expression according to

$$I(w,g) = I_{SD}(w) \cap I_{SR}(w) \cap I_{RD}(w)\,, \tag{92}$$

that shown in Fig. 5. *I(w,g)* represent the set good index for the relay channel polarized.

Now, at the last part of this section, the performance of polar codes for relay channels has been analyzed. The BER performance for this case has been shown when the relay-destination link is a Binary Symmetric Channel (BSC).

In Fig. 6 and Fig. 7, the performance of polar codes for DF and CF relaying in physically degraded relay channel has been shown, respectively. In this way, we utilize the proposed good index choosing based on the Section V. In this analysis, W_{SR} and W_{SD} are independent BSC. The Crossover probabilities for these links are equal to 0.05 and 0.15, respectively.

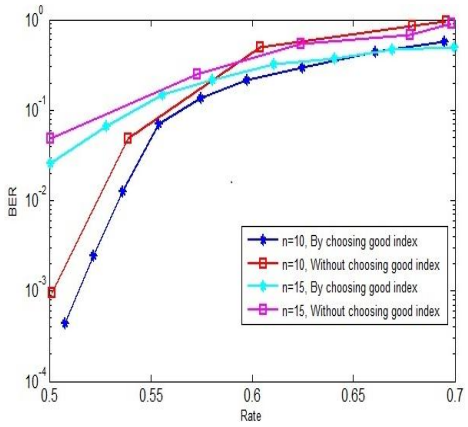


Fig. 6. Comparison of BER performance of polar codes for DF relaying for BSC with the proposed good index choosing method. It has been considered $I(W_{SR}) \approx 0.71, I(W_{RD}) \approx 0.53$ and $I(W_{SD}) \approx 0.31$ in the simulation.

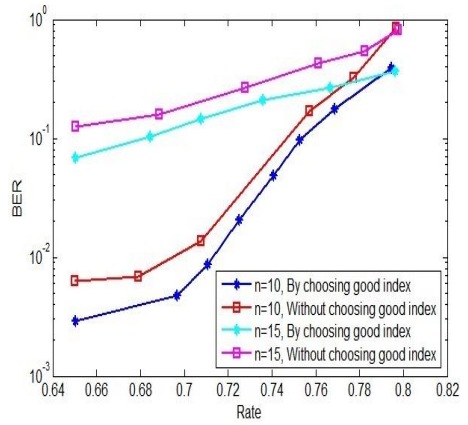


Fig. 7. Comparison of BER performance of polar codes for CF relaying for BSC with the proposed good index choosing method. It has been considered $I(W_{SR}) \approx 0.71, I(W_{RD}) \approx 0.53$ and $I(W_{SD}) \approx 0.31$ in the simulation.

By comparing Fig.6 and Fig.7, one can observe that using the proposed good index choosing method gives lower error probability.

6. conclision

In this paper, we showed that polar codes are suitable for DF and CF relaying with the orthogonal receiver and represent idea about channel polarization specifically for relay channel. It has been considered that for two relays, when the links are polarized, the capacity of one relay increases while the other decreases and the capacity region of the relay changes. It has been shown that polarization make improve the cut-set bound for relay channel. At last, we introduced a new scheme that shows how to choose a good index for polarized relay channels. By using the proposed method, one can find less error probability.References

References

[1] El Gamal, Abbas, and Young-Han Kim. Network information theory. Cambridge university press, 2011

[2] Cover, Thomas M., and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012.

[3] Arikan, Erdal. "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels."Information Theory, IEEE Transactions on 55, no. 7 (2009): 3051-3073.

[4] Arıkan, Erdal, and Emre Telatar. "On the rate of channel polarization." InInformation Theory, 2009. ISIT 2009. IEEE International Symposium on, pp. 1493-1495. IEEE, 2009.

[5] Sasoglu, Eren, Emre Telatar, and Edmund Yeh. "Polar codes for the two-user multiple-access channel." arXiv preprint arXiv:1006.4255 (2010).

[6] Abbe, Emmanuel, and Emre Telatar. "Polar codes for the-user multiple access channel." Information Theory, IEEE Transactions on 58, no. 8 (2012): 5437-5448.

[7] Tavakoli, Hassan. "Polarization of a Point-to-Point Channel by a Multiple Access Channel: A New Method for Different Channel Polarization." Iranian Journal of Science and Technology, Transactions of Electrical Engineering41, no. 2 (2017): 115-122.

[8] Goela, Naveen, Emmanuel Abbe, and Michael Gastpar. "Polar codes for broadcast channels." Information Theory, IEEE Transactions on 61, no. 2 (2015): 758-782.

[9] Mondelli, Marco, S. Hamed Hassani, Igal Sason, and Rudiger L. Urbanke. "Achieving Marton's region for broadcast channels using polar codes."Information Theory, IEEE Transactions on 61, no. 2 (2015): 783-800.

[10] Andersson, Mattias, Vishwambhar Rathi, Ragnar Thobaben, Jörg Kliewer, and Mikael Skoglund. "Nested polar codes for wiretap and relay channels."Communications Letters, IEEE 14, no. 8 (2010): 752-754.

[11] Blasco-Serrano, Ricardo, Ragnar Thobaben, Mattias Andersson, Vishwambhar Rathi, and Mikael Skoglund. "Polar codes for cooperative relaying." Communications, IEEE Transactions on 60, no. 11 (2012): 3263-3273.

[12] Blasco-Serrano, Ricardo, Ragnar Thobaben, Vishwambhar Rathi, and Mikael Skoglund. "Polar codes for compress-and-forward in binary relay channels." In Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on, pp. 1743-1747. IEEE, 2010.

[13] Karzand, Mohammad. "Polar codes for degraded relay channels." InProceedings of the International Zurich Seminar on Communications, pp. 59-62. 2012.
[14] Wang, Lele. "Polar coding for relay channels." In Information Theory (ISIT), 2015 IEEE International Symposium on, pp. 1532-1536. IEEE, 2015.
[15] Khalili, Ramin, and Kavé Salamatian. "On the achievability of cut-set bound for a class of erasure relay channels: The non-degraded case." InProceedings of the Intl. Symposium on Information Theory and its Applications, pp. 10-13. 2004.
[16] Şaşoğlu, Eren. "An entropy inequality for q-ary random variables and its application to channel polarization." In Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, pp. 1360-1363. IEEE, 2010.
[17] Şasoğlu, Eren, Emre Telatar, and Erdal Arikan. "Polarization for arbitrary discrete memoryless channels." In Information Theory Workshop, 2009. ITW 2009. IEEE, pp. 144-148. IEEE, 2009
[18] Hassan Tavakoli, "Capacity and Channel Coding for Wireless Point-to-Point Z2 Channel", Wireless Personal Communications, https://link.springer.com/article/10.1007%2Fs11277-017-4945-

Hassan Tavakoli was born in Tehran, Iran on April 4, 1983. He received MSc degrees, in 2007, and PhD degree, in 2012, all in Communication Engineering from K. N. Toosi University of Technology in Iran. He is with the University of Guilan, Rasht as a faculty member since 2013. His research interests include Secure Communications, Error Control Coding Schemes, Machine Learning and Opimizing Telecommunications Systems.

Saeid Pakravan received the B.Sc. degree in communication engineering from University of Birjand, Birjand, Iran, in 2014. He received the M.Sc. degree in communication engineering from Guilan University, Rasht, Iran, in 2017. His area resarch interests include Information Theory, Channel Coding and Wireless Communication. His email address is: Saeidpak70@yahoo.com