Asian Journal of Engineering and Applied Technology
ISSN: 2249-068X (P) Vol. 1 No. 1, 2012, pp.12-18
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajeat-2012.1.1.2506

Topic Oriented Probability Based and Semi Supervised Document Clustering

M. Karthikeyan and P. Aruna
Department of Computer Science and Engineering
Annamalai University, Annamalai Nagar - 608 002, Tamil Nadu, India
E-mail: mkshkarthik@yahoo.co.in

(Received on 20 December 2011 and accepted on 12 March 2012)

Abstract - **Clustering of related or similar objects has long been regarded as a potentially useful contribution for helping users to navigate an information space such as a document collection. But, the major challenge in document clustering is high dimensionality. Data mining and statistical techniques have been applied with some success to large set of documents to automatically produce meaningful subsets. Many clustering algorithms and techniques have been developed and implemented since the earliest days of computational information retrieval but as the sizes of document collections have grown, these techniques have not been scaled to large collections because of their computational overhead. Traditional document clustering is usually considered as an unsupervised learning. It cannot effectively group documents under user's need. To solve this problem, the proposed system concentrates on an interactive text clustering methodology, topic oriented probability based and semi supervised document clustering. It suggests interactive approach for document clustering, to facilitate human refinement of clustering outputs. The proposed system evaluates system efficiency by implementing and testing the clustering results with Dbscan and K-means clustering algorithms. Experiment shows that the proposed document clustering algorithm performs with an average efficiency of 94.4% for various document categories.**

Keywords : **Document Clustering, Text Documents, Word Frequency, Probability, Tokenization, Structural Filtering**

I. Introduction

With the rapid development of Information technology, the number of electronic documents and digital content of documents exceed the capacity of manual control and management. People are increasingly required to handle wide ranges of information from multiple sources. As a result, document clustering techniques are implemented by organizations to manage their information and knowledge more effectively. Document clustering can be defined as the task of learning method for categorizing electronic documents into their automatically annotated classes, based on its contents. It is widely applicable in areas such as search engines, web mining, information retrieval and topological analysis. Document clustering is a critical component of research in text mining. Traditional document clustering includes: (a) Extracting feature vector of a document, (b) Clustering document by parameters such as similarity threshold, the number of clusters etc. However, traditional document clustering uses an unsupervised learning. It cannot effectively group documents under need of user (Jiangtao Qiu, Changjie Tang, 2007). Although many clustering algorithms have been proposed in the literature, most of them do not satisfy the special requirements for clustering documents.

1. High Dimensionality: The number of relevant terms in a document set is typically in the order of thousands. Each of these terms constitutes a dimension in a document vector. Natural clusters usually do not exist in the full dimensional space, but in the subspace formed by a set of correlated dimensions. Locating clusters in subspaces can be challenging.

2. Scalability: Real world data sets may contain hundreds of or thousands of documents. Many clustering algorithms work fine on smaller data sets, but fail to handle large data sets efficiently.

3. Accuracy: A good clustering solution should have high intra-cluster similarity and low inter-cluster similarity, i.e. documents within the same cluster should be similar but are dissimilar to documents in other clusters.

4. Easy to Browse with Meaningful Cluster Description: The resulting topic hierarchy should provide a sensible structure, together with meaningful cluster descriptions, to support interactive browsing.

5. Prior Domain Knowledge: Many clustering algorithms require the user to specify some input parameters, e.g. the number of clusters. However, the user often does not have such prior domain knowledge. Clustering accuracy may degrade drastically if an algorithm is too sensitive to these input parameters.

Based on these observations, the proposed method concentrates on topic oriented probability based and semi supervised document clustering approach. Proposed system uses vector space model to represent the documents. A document in the vector space model is represented as a weight vector, in which each component weight (frequency) is computed based on some variation of term frequency scheme. In this method, the weight of a word ti in document dj is the number of times that t_i appears in document d_j, denoted by f_{ij}. Normalization of the word is done as per the equation 1. The shortcoming of the word frequency scheme is that it does not consider the situation where a word appears in many documents of the collection, such word may be discriminative.

$$\mathrm{tf}_{ij} = \frac{f_{ij}}{\max\ \{f_{1j}\ ,f_{2j}\ , \ldots, f_{|v|j}\}} \qquad (1)$$

Where the max is computed over all words that appear in document d_j. If a word does not appear in dj then tf_{ij} =0, |v| is the vocabulary size of the collection.

II. Related Works

Document clustering is a traditional subject in data mining and machine learning. In recent studies, many new technologies are introduced. Benjamin C.M. Fund, Ke Wang and Martin Ester [1] proposed the method for hierarchical document clustering using frequent data item sets. Chihli Huang, Stefan Wermter and Peter Smith [2] exploit WordNet's hyponym relationship to obtain fewer but more general concepts and thus further improve SOM's documents clustering ability. Some researchers use ontology to improve document clustering performance. Hotho A, Maedche A,Staab S [3] proposed Ontology-Based document clustering. Yuan-Chao Liu, Xiao-Long Wang, Bing-Quan Liu [4] presents a feature selection algorithm for document clustering based on word co-occurrence frequency. In this algorithm, the impact of feature selection on document clustering is discussed firstly, and then a new solution for feature selection was brought forward which is based on word co-occurrence frequency.

Zamir [5] proposed suffix tree to find the maximum word sequence (phases) between two documents. Bakus, Hussin, and Kamel [6] used a hierarchical phrase grammar extraction procedure to identify phrases as features for document clustering. The self organizing map (SOM) method was used as the clustering algorithm. Mladenic and Grobelink [7] used Naïve Bayesian method to classify documents based on word sequences of different length. Experimental results show that using the word sequences whose length is no more three words can improve the performance of text classification system. But when the average length of used word sequences is longer than three words, there will be no difference between using word sequences or single words.

Previous semi-supervised approaches fall into three categories: constraint-based, metric-based and the combined approaches. Constraint based approaches explicitly modify the objective function or make certain constraints during the clustering process. It is illustrated by K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl [8]. Metric based approaches parameterize distance metric and learn the metric parameters in a manner, so that the distance between objects connected by must-links is smaller and the distance between objects connected by cannot-links is larger in general. It is illustrated by Sugato Basu, Mikhail Bilenko, and Raymond J.Monney [9]. Both constraint-based and metric-based approaches are combined and used by Sugato Basu, Mikhail Bilenko, and Raymond J.Monney [10]. Another recent approach of incorporating prior knowledge tackles the problem differently and it is illustrated by David Gondek and Thomas Hofmann [11]. They defined the non-redundant data clustering as a problem of discovering alternate clustering solutions given a known clustering solution. Hsin-Chang Yang, Chung-Hong Lee [12] proposed a text mining approach for automatic construction of hypertext using SOM algorithm. Todsanai Chumwatana, Kok Wai Wong and Hong Xie [13] proposed non segmented document clustering method using self-organizing map (SOM) and frequent max substring technique to improve the efficiency of information retrieval.

Linghui Gong, Jianping Zeng, Shiyong Zhang [14] focuses on the problem of adaptive feature selection for clustering text stream. They proposed a validity index based method of adaptive feature selection, incorporating with which a new text stream clustering algorithm is developed. Wei Song, Cheng Hua Li, Soon Cheol Park [15] proposes a self-organized genetic algorithm for text clustering based on ontology method. They implemented two hybrid strategies using various similarity measures to investigate how ontology methods could be used effectively in text clustering. Ramiz M. Aliguliyev [16] shows that assignment of weight to documents improves clustering solution. Ridvan Saracoglu, Kemal Tutuncu, Novruz Allahverdi [17] proposed a new approach on search for similar documents with multiple categories using fuzzy clustering. Dino Isa, V.P. Kallimani, Lam Hong Lee [18] describes the implementation of an enhanced hybrid classification approach which affords better classification accuracy through the utilization of two familiar algorithms, the naïve bayes classification algorithm and the self organizing map clustering algorithm.

Ramiz M. Aliguliyev [19] proposed a new sentence similarity measure and sentence based extractive technique for automatic text summarization. Pei-Yi Hao, Jung-Hsien Chiang, Yi-Kun Tu [20] proposed a novel hierarchical classification method that generalizes support vector machine learning. Yuen-Hsien Tseng [21] described a cluster labeling algorithm for creating generic titles, based on external resources such as WordNet. Shih-Cheng Hong, Feng-Yi Yang, Shieh-Shing Lin [22] proposed a hierarchical fuzzy clustering decision tree for the classification problem with large number of classes and continuous attributes. Lam Hong Lee, Dino Isa [23] proposed a method for automatically computed document dependent weighting factor facility for Naïve Bayes classification. Chia-Hui Chang, Zhi-Kai Ding [24] introduced a new clustering approach, Categorical data clustering with subjective factors. They developed a visualization tool for clustered categorical data such that the result of adjusting parameters is instantly reflected.

III. System Overview

Topic oriented probability based and semi supervised document clustering is defined as follows: Given a set S of n documents and a set T of k topics, proposed system like to partition the documents into k subsets S1, S2, ….Sk, each corresponding to one of the topics, such that (i). Documents assigned to each subset are more similar to each other than the documents assigned to different subsets, and (ii). Documents of each subset are more similar to its corresponding topic than the rest of the topics. The functional components and data flow of proposed probability based topic oriented and semi supervised document clustering method is depicted in figure 1.

The major steps involved in the proposed method are given below:

1. Documents of various categories are collected and stored in the database.
2. All the words which are appearing in the documents are extracted and stored in a separate words database.
3. From the database, distinct words are identified and the probability of each word is calculated.
4. During the clustering process, based on the higher probability of words, the documents are classified and clustering of related documents is done.
5. The output of the proposed method are compared and verified for justification by using the famous data mining partition based clustering algorithm K-means and density based clustering algorithm Dbscan.
6. Finally, results are analyzed and compared.

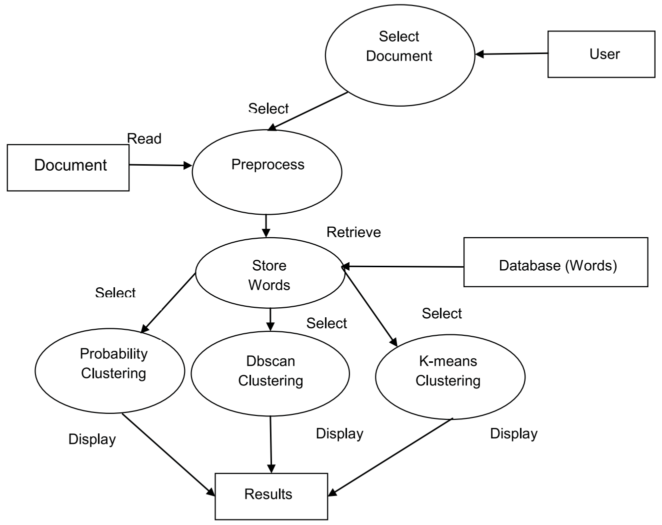

Fig. 1 The functional components and data flow of the proposed method

A. Document Clustering by Topic Oriented Probability Based and Semi Supervised Clustering Algorithm

The proposed new documents clustering method, group documents according to user's need. The main steps includes: (1) Design a multiple-attributes topic structure to represent user's need. (2) Make topic-semantic annotation for each document, and then compute topic-semantic similarity between documents. (3) Compute words probability, and (4) Group documents based on maximum word probability. The main objective is to reduce dimensionality of feature vector. Dimensionality reduction of feature vector is a hotspot of research in text mining. The dimensionality of document vectors may reach thousands and even tens of thousands. It results in huge time cost on documents clustering. Because only words that are able to being mapped to attributes of topic will be extracted from a document, the proposed methods reduce dimensionality effectively.

The proposed method is divided into 4 modules. They are:

1. Training
2. Probabilistic scanning
3. Dbscan clustering
4. *K*-means clustering

1. Training

Five different categories of documents are used for training purpose as a sample document corpus. They are Business documents, Education documents, Politics documents, Medicine documents and Sports documents. But it is possible to extend the categories. The aim is to derive higher level concepts from the words of the different categories of document corpus in order to populate the knowledge base (database). In the first task, words are extracted from the training documents and they are matched with other words, or with ones that already exist in the database. To achieve this goal, a pipeline of analysis that contains two stages is defined. The two stages of task include:

Tokenization: This is the very first linguistic analysis step. It consists of breaking the free text into a sequence of separate words and punctuation symbols (tokens). Its input consists of natural language text and the output contains a list of the tokens extracted.

Structural Filtering: This stage uses the output of the tokenization and keeps, discovers or discards words according to contextual information. The actual module is a rule compiler which applies filtering based on rules like, words length less than 3 characters and greater then 20 characters etc., Option is given in the proposed system to omit stop words like,"Can","are","has","with","the","they","which"," have",etc. Similarly, words above 15 or 20 characters are also omitted because these words will not be distinct words for document clustering process. Then, the filtered texts (words) are called distinct words and these are stored in the database based on selected category by the user. Then, the probability of occurrence is calculated for distinct words. For example given a set S of N documents and a set T of K topics, then the probability of distinct words will be calculated by εWi / Wj Where Wi denotes the number of occurrences of a distinct word in a training document and Wj denotes the total number of distinct words in a training document. Similarly for K topics of set T, the probabilities of occurrences of distinct words are calculated for a set S of N documents.

2. Probabilistic Scanning

The probabilistic scan works as classification of a document. It is done on the basis of comparing the probability of occurrence of distinct words in the database. Based on the probability, the count for each category is calculated. Proposed method calculates a probability for distinct words by measuring similarity between documents and evaluating clustering partitions. In this context, and more generally throughout information retrieval, a commonly used measure of similarity is obtained by representing documents as normalized vectors. Each dimension of the vector corresponds to a distinct word in the union of all words. A document is then represented as a vector containing the normalized frequency count of the words in it. Intuitively, this measure tries to capture the degree of word overlap between two documents.

Input : Set S of N documents and a set T of K topics

Output : Array of distinct words and array of distinct words count based on K

Method :

```
1.  Read S, N ; // Read the training documents one by one to split all words
2.  Read T, K; // Read the topics
3.  Preprocess D to get $W_i$ // Preprocess to identify distinct words
4.  For each $W_i$ in D
5.  For each K of T
6.  $P_i = \varepsilon W_i / W_j$ // Probability calculation of distinct words
7.  Count = count+1 // Based on the topic count will be calculated
8.  End
9.  End
10. Return $w_i$
11. Return count
```

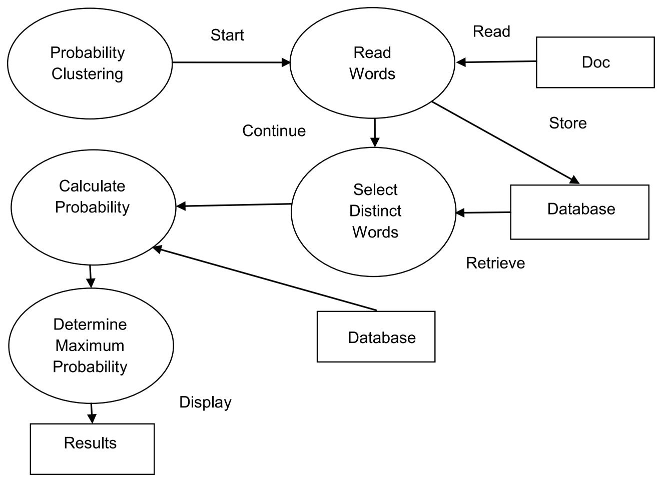

Fig. 2 Functional components and data flow of proposed topic oriented probability based and semi supervised document clustering method

3. Testing

The proposed probability based document clustering method is compared with existing Dbscan and K-means clustering algorithms. K-means is a partition clustering algorithm based on iterative relocation that partitions a dataset into k clusters. The Objective function locally minimizes sum of squared distance between the data points and their corresponding cluster centers to form the clusters. Figure 3 shows how clustering results are tested by K-means algorithm. The key idea of the Dbscan algorithm is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. This algorithm needs three input parameters: k, the neighbor list size, Eps, the radius that delimitate the neighborhood area of a point (Eps neighborhood), MinPts, the minimum number of points that must exist in the Eps-neighborhood. The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points (directly density-reachable, density-reachable, density-connected) to form the clusters. Figure 3 shows how clustering results are tested by Dbscan algorithm.

4. Efficiency

For comparing the proposed method with the existing algorithms Dbscan and K-means, efficiency is calculated by using the formula,

$$\frac{\text{Number of documents classifed correctly}}{\text{Total number of documents stored in database}}$$

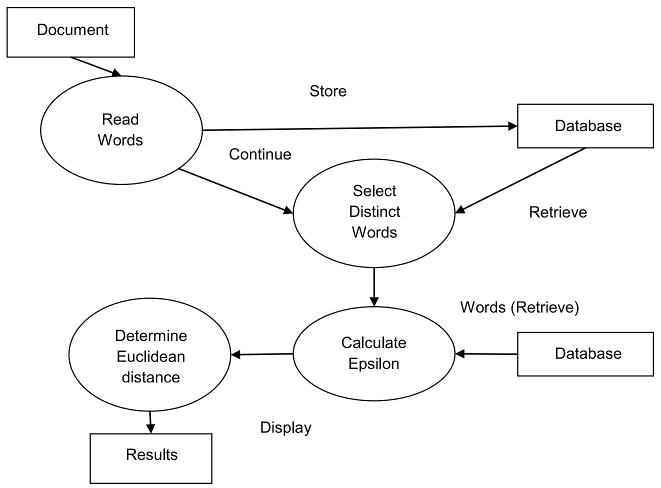

Fig. 3 Functional components and work flow of DBSCAN Algorithm for document clustering

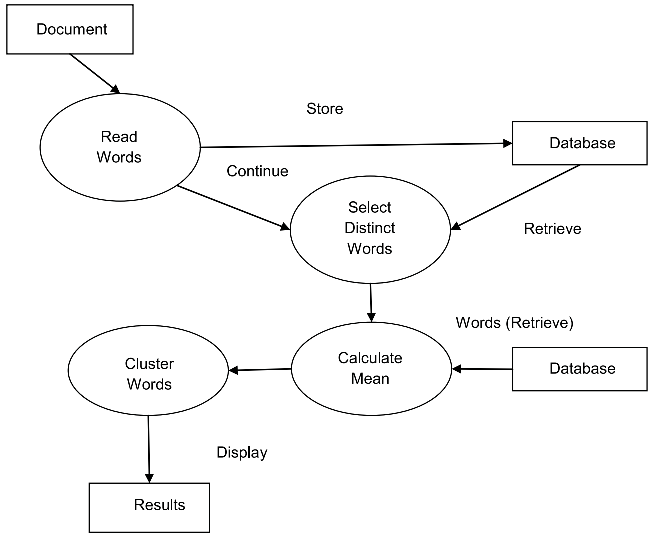

Fig. 4 Functional components and work flow of K-means Algorithm for document clustering

IV. Experimental Results And Discussions

In experiments, totally 2000 documents of five different categories are used for training purpose. The system was trained according to the category selected by the user. The categories includes: 1. Business documents 2. Education documents 3. Political documents 4. Medicine documents 5. Sports documents. Proposed system omits nearly 500 stop words. The stop words in document corpus are removed before distinct words extraction. Then, distinct words are extracted and their frequency of occurrence is calculated. Based on the frequency, probability is calculated. Table 1 shows training process results of topic oriented probability based and semi supervised document clustering algorithm.

Totally during experiments 3005794 distinct words are identified. According to the category selected by the user, the probability was calculated. Based on the probability of distinct words clustering is done during the testing phase. For testing, 600 documents (125 documents in each category) are used. Table II shows the accuracy achieved by topic oriented probability based and semi supervised document clustering algorithm for each category of the documents. The same documents are used as test case for Dbscan and K-means clustering algorithms respectively. Table III shows accuracy achieved by Dbscan algorithm and Table IV shows the accuracy achieved by the K-means algorithm. When comparing the existing algorithms Dbscan and K-means algorithms, the proposed probability based topic oriented and semi supervised document clustering algorithm yields better results. Figure 5 shows overall performance comparison of all the three clustering algorithms. The Dbscan algorithm performs with an average efficiency of 93.52% for all categories. K-means algorithm performs with an average efficiency of 91.68% for all categories. The proposed document clustering algorithm performs with an average efficiency of 94.4% for all categories of documents. The result shows that the proposed topic oriented probability based and semi supervised document clustering algorithm outperforms other two existing algorithms.

Table 1 Training Process Results of Topic Oriented Probability Based and Semi Supervised Document Clustering Algorithm

Sl.No.	Category	Number of documents Used for Training	Total Number of Distinct Words Extracted
1	Business	400	630836
2	Education	400	652416
3	Politics	400	548010
4	Medicine	400	562046
5	Sports	400	612486

Table II Accuracy of Probability Based Topic Oriented and Semi Supervised Document Clustering Algorithm

Category	Number of Documents Used for Testing	Number of Documents Classified Correctly	Percentage
Business	125	123	98.40%
Education	125	122	97.60%
Politics	125	118	94.40%
Medicine	125	115	92.00%
Sports	125	112	89.60%

Table III Accuracy Of Dbscan Algorithm

Category	Number of Documents Used for Testing	Number of Documents Classified Correctly	Percentage
Business	125	120	96.00%
Education	125	119	95.20%
Politics	125	114	91.20%
Medicine	125	110	98.00%
Sports	125	109	87.20%

Table IV Accuracy of K-Means Algorithm

Category	Number of Documents Used for Testing	Number of Documents Classified Correctly	Percentage
Business	125	121	96.80%
Education	125	116	92.80%
Politics	125	115	92.00%
Medicine	125	111	88.80%
Sports	125	110	88.00%

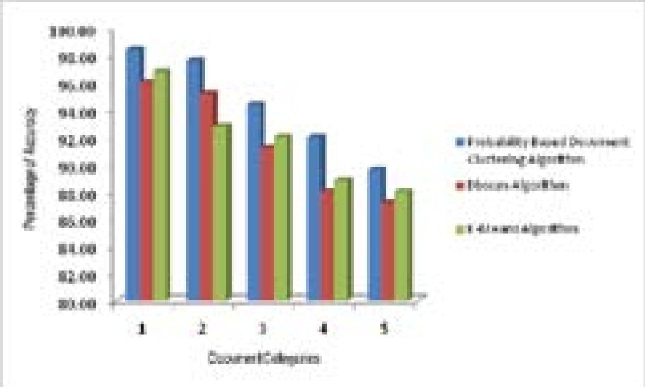

Fig. 5 Overall performance comparison of all the three clustering algorithms

VI. Conclusion

Traditional document clustering are unsupervised learning approaches. Traditional approaches often fail to obtain good clustering solution when users want to group documents according to their need. Focusing on this problem, the proposed method uses the topic oriented and probability based document clustering to fulfill the user requirement. Further the proposed method was compared and checked with the famous clustering algorithms Dbscan and K-means. In future probability similarity score to include arbitrary functions over words in documents (such as phrases and logical operations) may be implemented. This can be done by expanding the domain of the multi-nominal distributions currently used to compute expected document overlap and similarity measure may be extended to problems in other domains such as video segmentation, audio classification using different estimation techniques as appropriate. Although two thousand documents of various categories used for training purpose in the proposed system, it is scalable according to the user's need. Experiments show that the proposed method is feasible and effective.

References

[1] Benjamin C.M. Fung, Ke Wang and Martin Ester. "Hiearchical Document Clustering Using Frequent Item sets". *Proceedings of the SIAM International Conference on Data Mining,* 2003.

[2] Chihli Huang, Stefan Wermter and Peter Smith, "Hybrid Neural Document Clustering Using Guided Self-Oranization and WordNet". *Intelligent Systems,* IEEE,3/2004.

[3] Hotho A, Maedche A,Staab S. "Ontology-Based document clustering". *In Proc. Of the workshop Text Learning: Beyond Supervision,* at IJCAI 2001. Seattle, WA,USA, Aug 6.

[4] Yuan-Chao Liu, Xiao-Long Wang, Bing-Quan Liu, "A Feature Selection Algorithm for Document Clustering Based on Word Co-occurrence frequency" at *Third International Conference on Machine Learning and Gybernetics,* Shanghai,26-29 Aug. 2004.

[5] O. Zamir, Clustering Web Documents: "A Phrase-Based Method for Group Search Engine Results", Ph.D. dissertation, Dept. Computer Science & Engineering, Univ. of Washington, 1999.

[6] Bakus, Hussin, and Kamel," A SOM-Based Document clustering Using Phrases", In *proceeding of the 9th International Conference on Neural Information Processing,* Vol. 5,2002,pp 2212-2216.

[7] D. Mladenic and M. Grobelink, " Word Sequence as Features in Text-learning", *In proceedings of the 17th Electrotechnical and Computer Science Conference,* Ljublijana, Slovenia, 1998.

[8] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained k-means clustering with background knowledge. *In proc. Of 18th International Conference on Machine Learning,* pp. 577-584, 2001.

[9] Sugato Basu, Mikhail Bilenko, and Raymond J.Monney, "A probabilistic framework for semi-supervised clustering". *In Proc. Of the 10th International Conference on Knowledge Discovery and Data Mining* 2004.

[10] Mikhail Bilenko, Sugato Basu and Raymond J. Monney, " Integrating constraints and metric learning in semi-supervised clustering". In *Proc. Of 21st International Conference on Maching Learning,* 2004.

[11] David Gondek and Thomas Hofmann, "Non-redundant data clustering". *In proc. Of the fourth IEEE International Conference on Data Mining,* 2004.

[12] Hsin-Chang Yang, Chung-Hong Lee, "A text mining approach for automatic construction of hypertexts", *Expert Systems with Applications,* Vol. 29, 2005, pp. 723-734.

[13] Todsanai Chumwatana, Kok Wai Wong and Hong Xie, " A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts", *J. Intelligent Learning Systems & Applications,* Vol. 2, pp. 117-125, 2010.

[14] Linghui Gong, Jianping Zeng, Shiyong Zhang, " Text stream clustering algorithm based on adaptive feature selection", *Expert Systems with Applications,* Vol.38, No. 3, 2011, pp.1393-1399.

[15] Wei Song, Cheng Hua Li, Soon Cheol Park, " Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures", *Expert Systems with Applications,* Vol.36, No. 5, 2009, pp. 9095-9104.

[16] Ramiz M. Aliguliyev, " Clustering of document collection – A weighting approach", *Expert Systems with Applications,* Vol.36, No.4, 2009, pp. 7904-7916.

[17]. Ridvan Saracoglu, Kemal Tutuncu, Novruz Allahverdi, " A new approach on search for similar documents with multiple categories using fuzzy clustering", *Expert Systems with Applications*, Vol.34, No. 4, 2008, pp. 2545-2554.

[18] Dino Isa, V.P. Kallimani, Lam Hong Lee, " Using the sellf organizing map for clustering text documents", *Expert Systems with Applications,* Vol.36, No. 5, 2009, pp. 9584-9591.

[19] Ramiz M. Aliguliyev, " A new sentence similarity measure and sentence based extrative technique for automatic text summarization", *Expert Systems with Applications,* Vol.36, No. 4, 2009, pp.7764-7772.

[20] Pei-Yi Hao, Jung-Hsien Chiang, Yi-Kun Tu, " Hierarchically SVM classification based on support vector clustering method and its application to document categorization", *Expert Systems with Applications,* Vol.33, No. 3, 2007, pp. 627-635.

[21] Yuen-Hsien Tseng, "Generic title labeling for clustered documents", *Expert Systems with Applications*, Vol.37, No.3, 2010, pp. 2247-2254.

[22] Shih-Cheng Hong, Feng-Yi Yang, Shieh-Shing Lin, "Hierarchical fuzzy clustering decision tree for classifying recipes of ion implanter", *Expert Systems with Applications,* Vol.38, Issue 1 (2011) 933-940.

[23] Lam Hong Lee, Dino Isa, "Automatically computed document dependent wighting factor facility for Naïve Bayes classification", *Expert Systems with Applications,* Vol.37, No. 12, 2010, pp. 8471-8478.

[24] Chia-Hui Chang, Zhi-Kai Ding, "Categorical data visualization and clustering using subjective factors", *Data & Knowledge Engineering,* Vol. 53, 2005, pp. 243-262.