Recent Developments in Text Clustering Techniques

Saurabh Sharma
M.E. Research Scholar, CSE,
University Institute of Engineering & Technology,
Panjab University, Chandigarh, India

Vishal Gupta
Assistant Professor, CSE,
University Institute of Engineering & Technology,
Panjab University, Chandigarh, India

ABSTRACT
In order to make better business decisions, faster database browsing and reducing processing time of queries, Extraction of Information from text documents in efficient manner is needed. Clustering of huge number of text documents into different clusters, for better management of information, provides for a wide area in which a whole lot of research is currently being pursued. Recent developments in this area have tried number of different techniques. This paper reviews and discusses "Text Clustering" and partially covers all major techniques currently in use for the Process.

General Terms
Text Document Clustering, Text Clustering, document Clustering, Text Mining.

Keywords
Text clustering, K-mean clustering, hierarchical clustering, topic tracing, feature selection, ontology, WORDNET, frequent word sequence.

1. INTRODUCTION
Text document clustering is a text mining technique which divides the given set of text documents into significant clusters. It is used for condensing a huge number of text documents into well-organized form for getting the desired results in less time. In recent advancements in this field, use of Word Net like additional knowledge resources has increased for the purpose of external knowledge base for better clustering results. Clustering is an effective method for search computing [1]. It offers the possibilities like: grouping similar results [2], comprehend the links between the results [3] and creating the succinct representation and display of search results.

To understand the concept of clustering, we take the example of a super market. In a super market, items for a variety of needs are available ranging from daily common needs to specific needs. These items are always placed in clusters or groups in different parts of super market. Similar kinds of items are placed near each other forming a type of cluster. For example, items related to kitchen are kept in one area and items related to electronics are placed in another area, and so on. To make search for items easier, items with same features or same type are placed together. And then each area is labeled with appropriate names. Now when a user wants a product of a particular item type, she would only have to go to that area and check for it instead checking all the areas of super market.

Text document clustering is mainly used for automatic detection and organization of text documents into meaningful groups based on some criteria chosen for that set of objects. Text document clustering uses keywords for information extraction process. Keywords are the important terms which describe the content to be extracted from documents. Different keywords are used for making different clusters from the same set of documents. So, we can say that selecting correct keyword is very important part of text document clustering for better and accurate results. With wrong keywords or content, poor keywords may lead to unexpected result even with the best techniques available. Text document clustering, in general, is considered a centralized process.

Normally, documents within a cluster are more similar to each other than documents lying in other clusters. The main objective of text document clustering is to generate such kind of clusters in which most similar documents fall into a single cluster i.e. maximize the similarity measure within single cluster documents and minimize the similarity measure among other cluster documents.[4][5][6][7][8]

It is important to understand the difference between clustering and classification. In classification, we are provided with a collection of pre-classified training data, and the problem is to label a newly encountered unlabeled data. Typically, the given labeled data (training data) is used to learn the descriptions of classes, which in turn are used to label new data. In the case of clustering, the problem is to group a given collection of unlabeled data into meaningful clusters. Since clustering needs little prior information about the data than classification, it could be applied in many situations where classification could not be performed.

Typical text clustering activity involves the following steps as shown in figure 1[4]:

- Document representation (optionally including feature extraction and/or selection).
- Definition of a document similarity measure.
- Clustering or grouping.

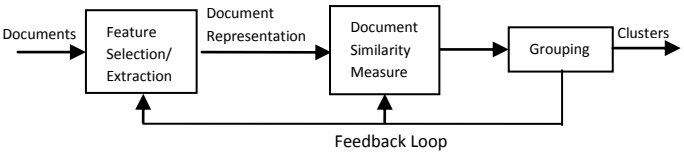

Figure 1: Stages in Text Clustering [9]

Fig. 1 shows a typical sequencing of these three steps, including a feedback path where the grouping process output could affect subsequent feature selection/extraction and similarity computations [5]. Document representation refers to the number of clusters, the number of documents, and the number, type and scale of the features available to the clustering algorithm. Feature selection is the process of identifying the most effective subset of the original features to use in clustering [4].

Some special requirements for text clustering are:

• Finding a suitable model to represent the document is a nontrivial issue. Most of the text documents are written in a human language, which is context-sensitive. As the accurate meaning of a sentence has a close relationship with the sequential occurrences of words in it, the document model better preserves the sequential relationship between words in the document.

• In the vector space model, the length of document vectors is normalized. However, in real life, the sizes of vocabularies for different topics are different from each other.

• In the real world, people may use different word forms to express the same word meaning, and the same word form to express different word meanings. In text clustering, if only word forms are used as features, the topics of documents may not be fully captured. Word meanings are better than word forms in terms of representing the topics of documents. Thus, it is beneficial to involve ontology into the text clustering algorithm.

• Not all the words in a document are closely related with the topic of the document. Documents without these irrelevant words will certainly provide valuable information for clustering. How to identify and remove them is a nontrivial problem.

• Associating a meaningful label to each final cluster is essential. Then, the user can easily find out what the cluster is about since the label can provide an adequate description of the cluster. However, it is time-consuming to determine the labels after the clustering process is finished.

• Overlapping between document clusters should be allowed because a document can cover multiple topics. For example, morning news may have information about a war followed by information about the popularity of teas in US.

• The high dimension of text documents should be reduced. Usually there are about 200-1000 unique words in a typical document. In order to efficiently process a huge text database like WWW, the text clustering algorithm should have a way to reduce the high dimension.

The number of clusters is unknown prior to the clustering. It is difficult to specify a reasonable number of clusters for a data set when you have little information about it. Instead of telling the clustering algorithm what is the number of clusters, it makes more sense to let the clustering algorithm find it out by itself [9].

Nearly every type of clustering techniques performs number of preprocessing steps which mainly include stop word removal and stemming on the text documents. After the preprocessing steps, each document, with remaining words, is represented in a form of vector with frequency of occurrence of words as a non negative integer value.

Few clustering techniques perform an additional preprocessing step in which occurrence frequency of a word is divided by the overall frequency of the word in the whole document set. This step is used to reduce the discriminating power of those words, which are very common in each document of the whole document set and hence no significance importance during clustering [10].

Although, so many techniques have been proposed in recent development of document clustering, but no one could satisfy these special requirements of clustering text documents:

High dimensionality: Every related term in a document is considered as a dimension for clustering. Since, each document may contain thousands of related terms; it is very hard to handle high dimensionality of this order for a document set containing thousands of documents. Number of meaningful clusters is very few which can actually be derived from the subspace formed by a set of correlated dimensions instead of processing all dimension, so there should be a mechanism for the dimensionality reduction. Locating clusters in subspaces can be challenging.

Scalability: In real world data sets, number of documents are normally lie in the range of few hundred to few thousands or more than that. Most of the clustering techniques could not process the huge size of data sets, which results in poor efficiency.

Accuracy: For getting the more accurate results, clustering technique should have high similarity among intra-cluster terms and low similarity among inter-cluster terms. i.e., documents within the same cluster should be similar but are dissimilar to documents in other clusters. An external evaluation method, the F-measure [9], is commonly used for examining the accuracy of a clustering algorithm.

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.

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 [11]..

2. TEXT DATA CLUSTERING TECHNIQUES

2.1 Vector Space Model

Vector space model is being used in most of the currently available document retrieval systems and in several text mining approaches. However it was originally introduced for indexing and information retrieval [12]. Documents are represented as vectors in m-dimensional space, i.e. a numerical feature vector can describe each document d

$$w(d) = (x(d;t1);:::;x(d;tm)) \qquad (1)$$

A document is considered as “bag of words”. The document is counted for the occurrence of each word. Each word forms a dimension in document vector. Thus, by use of simple vector operations the documents can be compared. Queries can be performed by encoding the query terms similar to the documents in a query vector. The query vector can then be compared to each document and by ordering the documents according to the computed similarity; a result list can be obtained [13].

To find an appropriate encoding of the feature vector is the main task of the vector space representation of documents. Each element of the vector usually represents a word (or a group of words) of the document collection, i.e. the size of the vector is defined by the number of words (or groups of words) of the complete document collection. Use of binary term vectors, i.e. a vector element is set to one if the corresponding word is used in the document and to zero if the word is not, is the simplest way of document encoding . If a query is encoded

in a vector, this encoding will result in a simple Boolean comparison or search. The importance of all terms for a specific query or comparison is considered as similar by using Boolean encoding

Usually term weighting schemes are used to improve the performance, where the weights reflect the importance of a word in a specific document of the considered collection. Large weights are assigned to terms that are used frequently in relevant documents but rarely in the whole document collection [14]. Thus a weight w(d; t) for a term t in document d is computed by term frequency tf(d; t) times inverse document frequency idf(t), which describes the term specificity within the document collection.[15].

2.2 Ontology

The bag of words representation ignores relationships between important terms that do not co-occur literally hence is often unsatisfactory for text clustering methods Meaningful sentences are composed of meaningful words; any system that aspires to process natural languages as people do, must have information about words and their meanings. This information is traditionally provided through dictionaries, especially, machine readable dictionaries, such as Word Net [16]. The traditional text representation method is based on the words which occur in the relative documents, and then the clustering methods compute the similarity between the vectors. However, many documents do not contain common words even though they contain the similar semantic information. For instance, if one document describes the "hockey" issue, it should be turned up "game" issue even though the document does not contain word "game". In order to deal with such problem, a concept-based model using ontologies is necessary [17][18].

Ontology has been defined by many and the most acceptable definition defines it as, according to Gruber [19]: "an ontology is a conceptual framework for defining the basic classes of entities in some domain of knowledge, the relationships these entities have to each other, and the organization of concepts in terms of higher-lever concepts, typically taxonomic in nature". The term ontology is often used to refer to a range of linguistic and conceptual resources, from thesauri and dictionaries containing records defining key terms, synonyms, and so forth in some domain, to taxonomies that classify names and phrases in higher level categories, and formal knowledge representations that might support automatic inferences and certain types of reasoning. According to the definition, ontology is a collection of concepts and their interrelationships which can collectively provide an abstract view of an application domain [20][21].

2.3 Hierarchical methods:

These Hierarchical methods of clustering documents, creates clustering classes in form of a tree structure which is called as dendogram.

This tree can be structured in two ways.

1) Start with considering each document into its own class. Then, two most similar documents are grouped into one class, this process continues until stopping criteria matched or number of user specified iterations completed. In the end, we get all documents combined into number of clusters.

2) Start with considering all documents belong to same class and then divide the class into two sub classes and so on till a stopping criteria matched or number of user specified iterations completed. The distance between two documents is used to find similarity between them. In the end, we get all the documents divided into number of clusters [22].

2.4 Partitioning Algorithms

One of the most common clustering algorithms is K-Means. There are many variants and each step can be elaborated on with different outcomes. These have been named differently.

K-Means Algorithm

1. Pick k objects at random and let them define k clusters.

2. Calculate cluster representatives.

3. Make new clusters, one per cluster representative. Let each text belong to the cluster with the most similar cluster representative.

4. Repeat from 2 until a stopping criterion is reached.

Random initial partition is defined by the first partition. The result depends on one of the many prevalent ways used for constructing. The centroid or the mean of the objects in the cluster is usually used as cluster representative. The cluster can be represented by the median or a few specific objects. The cluster representative may be calculated taking into consideration the fact that fuzzy objects may belong to several clusters and the clustering is fuzzy.

When no objects change clusters, or when there are a very few change clusters between iterations normally this is taken as stopping criterion. Since most quality improvement usually is gained during the first iterations, It may also be made to stop after a predefined number of iterations. Internal Quality Measure can be used to define the stopping criterion. The time complexity of the K-Means algorithm is O(knI), where k is the number of clusters, n the number of objects and I the number of iterations (which is dependent on the stopping criterion). The cluster representatives and the kn similarities between all objects and all clusters must be computed in each iteration [23].

A number of clusters is required by the K-Means algorithm as input. It implies that an appropriate number needs to be guessed. Clustering with the best result is reported, however, it is possible to run the algorithm with several different numbers of clusters to obtain the best results. Theoretically the result has the optimal number of clusters and in a general partitioning algorithm both splitting and division of clusters is allowed. Partitioning clustering may be viewed as an optimization problem. A Particular clustering setting a set of objects, a representation with a similarity measure and a number of clusters is termed as an instance. Clustering in this setting is an assignment.

The goal is to find a clustering with an optimal value which is returned by the function as a value for all clustering is it an objective, or criterion. An exhaustive search would be required to find such a clustering and local search strategies that are only guaranteed to find a local optimum is used by most partitioning clustering algorithms [24].

2.5 Document Clustering based on Topic Maps (TMHC)

Features like: words, phrases, and sequences from the documents are commonly used to perform clustering by document clustering algorithms [25][26][27][28]. To decide about the relatedness among documents, Simple features extraction techniques that are mainly based on feature counting and frequency distribution of the features are generally used by these algorithms. However, to cater to the

meaning behind the text (words) all these approaches have been insufficient. Clustering is done independent of the context by these techniques. Human language uses a context when writing Documents. The usage of words is largely dependent on this context. A standard for describing knowledge structures and later using it to support the find ability of that coded knowledge is being done by Topic Maps [29] .

To develop a vis-à-vis relation among the knowledge contents is the main emphasis of topic maps structures. Back -of-book indexes are merged for the creation of topic maps. This is the major operational area of the Topic Maps. Glossaries, Cross-references, Thesauri, or Catalogs use it as an effective tool for merging information from both structured and unstructured form. Set of assertions about one or more subjects are represented by a topic map. Three kinds of assertions are prevalent in use, namely, topic names, occurrences and associations. Topic maps for document clustering task have the major benefits which are as follows

(1) It helps reduce the size of document

(2) It can capture in a structured form, the topic related information from the document

(3) The inherit nature of arbitrary and robust information merging and

(4) It can easily handle the semantic topics and its hidden relationship and associations.

First step involves converting each document in a compact form which along with the occurrence and association between topics also represents the topics presented in the document. Second Step uses external source to generate topic maps information supporting various data representations. Input takes plain text files and returns topic maps as output for generating topic maps based on the information present in input text files.

A similarity measure is employed after collection of Topic Maps, which extracts the useful information; this information is used in three or more levels as the criterion for clustering. First the major topics are assigned to a document, secondly the tags are assigned that represent informational terms and after that the actual values corresponding to these tags. Relevant topics, tags and their values are then extracted. To document similarity matrix using this measure, another document is developed. In the Final step, Hierarchical agglomerative clustering is used to obtain the final clusters. Figure 2: shows the steps involved in topic maps based document clustering approach (TMHC) [30].

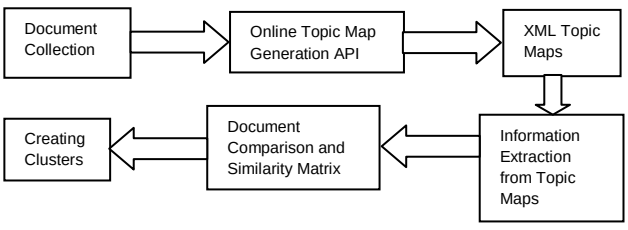

Figure 2: Steps involved in TMHC [30]

2.6 Document Clustering using Frequent Itemsets

The For Text Clustering, Frequently associated terms mined from large unstructured data source can be used. It can be done because frequent itemsets that are mined are able to drastically reduce the dimensionality of the documents.

A technique was developed for text clustering so that significant dimensionality reduction could be achieved in consonance with the frequent itemsets .To achieve this, a well-known method for mining the frequent itemsets, Apriori algorithm, is used. These frequent itemsets are then used to obtain, a set of non-overlapping partitions. Within the partition for the document collections, the resultant clusters are generated. A real life text dataset is used to carry out an extensive analysis of frequent item-based text clustering approach. Precision, Recall and F-measure were used to evaluate the Performance of the System which were fed on an input of the experimental results of the frequent item-based text clustering approach for 100 documents of dataset

The results ensured that the performance of the proposed approach improved effectively [8].

The devised approach consists of the following major steps:

(1) Text Preprocessing

(2) Mining of Frequent Itemsets

(3)Partitioning the text documents based on frequent itemsets

(4)Clustering of text documents within the partition

Algorithm: Text Clustering Process using frequent itemsets.

1. Collect the set of documents i.e. D = {d1, d2, d3, . dn} to make clusters.

2. Apply the Text preprocessing method on D.

3. Create the Binary database B.

4. Mine the Frequent Itemsets using Apriori algorithm on B.

5. Organize the output of first stage of Apriori in sets of frequent Itemsets of different length.

6. Partition the text documents based on Frequent Itemsets.

7. Cluster text documents within the zone based on their rank.

8. Output the resultant clusters.

2.7 Clustering Based on Frequent Word Sequences (CFWS) Algorithm

CFWS algorithm works in two steps: in step one, find sequence of words from each document and then group clustering candidates to have the final clustering.

In this algorithm, a text document d is analyzed and considered as word sequence, and then it can be represented as d= d =< w1;w2;w3; : : : >, where w1;w2;w3; : : : are words appearing in document d.

When support of a word sequence is equal to or greater than user specified threshold support, a word set is considered as frequent. This implies that this word set is available in at least the minimum number of documents specified by the user. A frequent k-word set is a frequent word set containing k words.

A word sequence is an ordered sequence of two or more words. A word sequence S is represented as < w1;w2;w3;w4 >, in which w1 is not compulsorily being followed by w2 in a text document.

It is pertinent to mention here that w1 is preceded by w2. However, words could intersperse between w1 and w2 which will not be included into the category of frequent words.

if these four words (w1;w2;w3, and w4) appear in the specified order in document d then A text document d supports this word sequence.

If a sequence occurs more than one time in the same document, such occurrence is counted as one.

Frequent Word Sequences can be ascertained by following two steps:-

(i) Finding frequent 2-word sets

(ii) Finding frequent word sequences of all length.

After ascertaining the Frequent Word Sequences, Clusters are created using word sequences [9].

2.8 Clustering Based on Frequent Word Meaning Sequences (CFWMS)

A word meaning [9] refers to the lexicalized concept that a word form can be used to express [31].

A single "Word Meaning" can be expressed by making use of different word form, which is called synonyms.

Different word forms which are synonyms to each other collectively called as synonym set, or shortly synset, is used to represent a word meaning.

Clustering result may be affected by lexical relation between word forms. For example, "movie" is a synonym of "film", so they can be used in place of the word and are exchangeable in documents. If the minimum support count of a word meaning is 10, and if 5 documents refer to movies by using “film” and 6 other documents use “movie” then sum of support count of these two word forms is greater than minimum support count. Now, if we consider these two word forms as synset for one word meaning, then these 11 documents can be placed into one cluster.

Text Clustering follows a very important context of Semantic relationship between word & meaning which is termed as Hyponymy/hypernymy. For example, the hypernym of a synset {pen, Pencil} is {writing material}, and the hypernym of {writing material} is {stationery}. The topics that are covered in articles under review may refer to the same information under Pen, Pencil or Writing Material/Stationery. Referring to one word as all encompassing, will limit the search and will not be exhaustive in approach. We may loose out on important information contained in and referred to by any other nomenclature.

Topics of documents are better described with word meanings than word forms. In CFWMS, Conversion of word meanings expressed by the word forms being used is done. After the conversion, each text document that will be analyzed is treated as a sequence of word meanings. If a word meaning sequence is found in a number of documents more than specified threshold, then it is marked as frequent word meaning sequence otherwise ignore it [9][31].

2.9 Clustering Based on Feature Selection

High dimensionality of attributes is major issue in performance of clustering algorithms. The major problem with high dimensionality is natural inbuilt sparseness of data in document set. Another problem is that very few attributes are important for clustering; many of them may result in less accurate results or even produce totally incorrect results, especially there are more unrelated features than related ones. [32]

Feature selection means selecting the key attributes for clustering based on class information available. Feature selection result in reduction from high dimensionality attributes to less number of significant attributes. This increases the overall performance and accuracy of the clustering algorithm. Feature selection also increases the understanding about data. The selected feature set should present most relevant information about the whole data set, needed for better clustering result [32]. Feature selection can improve the efficiency band accuracy of text classification algorithms by removing unnecessary and unrelated terms from the corpus [33]. Feature selection is very valuable dimensionality reduction method which is a crucial pre-processing technique to eliminate noisy features. Entropy measure gives better results for selecting the most relevant attributes because it did not affected by number of attributes, and it only change by the quality of clustering.

Widespread performance evaluation over synthetic, benchmark, and real datasets show its usefulness [34]. In iterative feature selection method clustering are iteratively performs and feature selection in a unified framework. Feature selection algorithms typically fall into two categories: feature ranking and subset selection [35]. Feature ranking ranks the features by a metric and removes all features that do not achieve a threshold value. Subset selection searches the set of expected features for the best possible subset [36].

3. CONCLUSION

Clustering of huge, diverse and rapidly changing text documents is very complex task. Clustering result mainly depends on the document set on which clustering is applied and parameters used for clustering criteria. For getting better clustering result, it is very important to select clustering parameters very precisely. Clustering has been gained much attention in last few years but more research is still needed for some issues. These include the achievement of better quality-complexity tradeoffs, as well as effort to deal with each method’s disadvantages. In addition, one more very significant issue is dimensionality, because the text documents may contain huge amount of terms. Another important issue is, often a document belongs to more than one cluster, this type of problem is referred to as an "any-of" problem. To deal with this issue algorithms should be developed which allow overlapping of clusters. In the end, additional efforts should be done to improvise the description of clusters contents from user’s point of view; this can be done with proper labeling and providing detailed information for each cluster.

4. REFERENCES

[1] Campi, A. and Ronchi, S., "The Role of Clustering in Search Computing ," in 20th International Workshop on Databases and Expert Systems Application , Linz, Austria, pp. 432-436, 2009. DOI: 10.1109/DEXA.2009.89

[2] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", in Fifteenth Annual International ACM SIGIR Conference, pp. 318-329, June 1992.

[3] Hearst, M. A. and Pedersen, J. O., "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," in 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 74-84,1996.

[4] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, Englewood Cliffs,1988.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323,1999.

[6] Congnan Luo, Yanjun Li, Soon M. Chung, "Text document Clustering Based on Neighbors", Data & Knowledge Engineering, Vol: 68, No: 11, pp: 1271-1288, November 2009.

[7] Xiangwei Liu, Pilian, "A Study On Text Clustering Algorithms Based On Frequent Term Sets", Advanced Data Mining and Applications, Lecture Notes in Computer Science, 2005, Vol. 3584/2005, pp. 347-354, DOI: 10.1007/11527503_42.

[8] S. Suneetha, Dr. M. Usha Rani, Yaswanth Kumar.Avulapati, "Text Clustering Based on Frequent Items Using Zoning and Ranking", International Journal of Computer Science and Information Security, Vol. 9, No. 6, pp. 208-209, June 2011

[9] Yanjun Li, "High Performance Text Document Clustering" Wright State University, 2007.

[10] Van Rijsbergen, C. J., "Information Retrieval", London: Butterworth Ltd., second edition.1979.

[11] Benjamin C. M. Fung, Ke Wang, and Martin Ester, "Hierarchical Document Clustering", Encyclopedia of Data Warehousing and Mining, pp. 555-559, 2005, DOI: 10.4018/978-1-59140-557-3.ch105

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing", Communications of the ACM, 18(11): pp. 613–620, 1975. (see also TR74-218, Cornell University, NY, USA)

[13] G. Salton, J. Allan, and C. Buckley, "Automatic structuring and retrieval of large text files", Communications of the ACM, 37(2): pp. 97–108, Feb 1994.

[14] G. Miller, "Wordnet: A Lexical Database for English," CACM, vol. 38, no. 11, pp.39-41, 1995.

[15] Andreas Hotho, Andreas N¨urnberger, Gerhard Paaß, "A Brief Survey of Text Mining", Journal for Computational Linguistics and Language Technology, pp. 27, 2005

[16] L. Khan, "Ontology-based Information Selection," PhD Thesis, 2000.

[17] L. Khan and D. McLeod, "Audio Structuring and Personalized Retrieval Using Ontology," Proceedings of IEEE Advances in Digital Libraries, 2000.

[18] T. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[19] Thomas R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", Proceedings of International Workshop on Formal Ontology, 1993.

[20] Liping Jing, "Survey of Text Clustering", The University of Hong Kong, HongKong, China, pp.3-4, 2005

[21] Abdelmalek Amine, Zakaria Elberrichi, and Michel Simonet, "Evaluation of Text Clustering Methods Using WordNet", International Arab Journal of Information Technology, Vol. 7, No. 4, pp. 351, October 2010

[22] D. J. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining", MIT Press, Cambridge, MA, USA. 2001 ISBN 0-262-08290-X.

[23] Magnus Rosell, "Introduction to Text Clustering", KTH CSC, pp. 14-15, September, 2008.

[24] Hammouda, K.M. and Kamel, M.S., "Efficient Phrase-Based Document Indexing for Web Document Clustering," IEEE Transaction on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1279-1296, 2004.

[25] Hung, C. and Xiaotie, D., "Efficient Phrase-Based Document Similarity for Clustering," IEEE Transaction on Knowledge and Data Engineering, vol. 20, no. September, pp. 1217-1229, 2008.

[26] Fung, B.C.M., Wang, K., and Ester, M., "Hierarchical Document Clustering Using Frequent Itemsets," Proceedings of SIAM International Conference on Data Mining, 2003.

[27] Soon, M. C. , John, D. H., and Yanjun, L., "Text Document Clustering Based on Frequent Word Meaning Sequences," Data& Knowledge Engineering, ELSEVIER vol. 64, pp. 381-404, 2008.

[28] Pepper, S., "Topic Maps," Encyclopedia of Library and Information Sciences, Third Edition 2010

[29] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq, "Document Clustering Based on Topic Maps", International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, pp. 33, December 2010

[30] C. Fellbaum (Ed.), "WordNet: An Electronic Lexical Database", MIT Press, May, 1998.

[31] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, March 2002

[32] Yanjun Li, Congnan Luo," Text Clustering with Feature Selection by Using Statistical Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 20 No.5, May 2008

[33] Manoranjan Dash ,Kiseok Choi ,Peter Scheuermann ,Huan Liu," Feature Selection for Clustering – A Filter Solution" Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)0-7695-1754-4/02 © 2002 IEEE

[34] Tao Liu, Shengping Liu , Zheng Chen, Wei-Ying Ma,"An Evaluation on Feature Selection for Text Clustering", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[35] MS. K.Mugunthadevi, MRS. S.C. Punitha, Dr..M. Punithavalli, "Survey on Feature Selection in Document Clustering" International Journal on Computer Science and Engineering, Vol. 3 No. 3, pp.1240-1241, Mar 2011

[36] Nora Oikonomakou and Michalis Vazirgiannis, "A Review of Web Document Clustering Approaches", Data Mining and Knowedge Discovery Handbook, VI, pp. 921-943, 2005, DOI: 10.1007/0-387-25465-X_43