Survey of Text Mining Techniques, Challenges and their Applications

N. Venkata Sailaja
Assistant Professor
Dept. of C.S.E
VNR VignanaJyothi Institute of Engg and Technology, Hyderabad, India

L. Padmasree, PhD
Professor
Dept. of E.C.E
VNR VignanaJyothi Institute of Engg and Technology, Hyderabad, India

N. Mangathayaru, PhD
Professor
Dept. of IT
VNR VignanaJyothi Institute of Engg and Technology, Hyderabad, India

ABSTRACT
In our everyday life communication interaction among people leading to mutual learning and sharing of valuable knowledge, such as chat, messaging, comments, and posts on board etc. Also, social networking websites, search engines sharing huge data texts in websites. The text is nothing but the combination of characters. Therefore, analyzing and extracting information patterns from such data sets are more complex. Several methods have been proposed for analyzing such texts and extracting information.

In this paper, we present different text mining techniques to discover various textual patterns from the different sources. This topic is also deals with the areas such as information retrieval, machine learning, statistics, computational data sciences and advanced data mining. We also discuss future challenges of this area using different techniques, particularly rough set based text mining techniques, improvements and research directions in this paper.

Keywords
Data mining, Text mining, Rough sets,Classification, Summarization, and Text categorization.

1. INTRODUCTION
Today's scenario in computer networks has become an essential of science and technology, so vast quantities of machine readable text documents become available. Today there is a lot of business information which exists in the form of text. Text mining allows us to deal with the inconsistent raw information, unstructured information and on the other hand allow handling with incomplete information, vagueness, uncertainty, and fuzziness.

A substantial piece of business information is put away in printed records accessible inside the web or intranets. The challenge is to extract that significant data and speak to its earning in a structure that can be effectively comprehended and reused by individuals or applications. In this manner the field of learning mining has been quickly extending, and drawing in numerous new scientists and clients. The fundamental explanation behind such a quick development is an incredible requirement for frameworks that can naturally

conclude new learning from the current data that is put away whether in organized or unstructured structures, from endless volumes of PC information being aggregated around the world.

The fields of data mining and text mining offer a guarantee for addressing to this need. The previous deal with data that is stored in an organized way in a database framework. The last deals with data that is stored in unstructured way inside the text documents.

1.1 Related work
Text mining techniques become more complex as compared to data mining due to unstructured and fuzzy nature of natural language text [19]. The techniquespresented in [13,16,21] comprises of multidisciplinary fields, such as information retrieval, text analysis, natural language processing [15], and information classification based on logical and non-trivial patterns from large data sets. In [17] Liu et al defined text mining as an extension of data mining technique. The data mining [18, 3] techniques are mainly used for the extraction of logical patterns from structured database. Text mining [14, 1] is similar to data mining, except that data mining tools [15] are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents. Text mining [8,16] is a knowledge discovery technique that provides computational intelligence.

The work of [7] demonstrated the possibility of accurately identifying the emotional tone of individual wordsin a corpus.Dhenakaran et al. [9,10] presented semantic web mining and its critical review where they addressed the problem of web retrieval systems (WIR) by providing machine interpretable semantics to provide greater machine support for the user. Stummea et al. [11] discussed the efficient results of web mining by exploiting semantic structures in the web, and using web mining techniques to build the Semantic Web. Aufaure et al. [12] showed the success of the Semantic Web depends [22] on the easy creation, integration and use of semantic data. Sampson et al. [13] pointed out that the Semantic web is the emerging landscape of new web technologies aimed at web-based information and services that would be understandable and reusable by both the humans and machines.

1.2 Organization of the paper
The rest of the paper is organized as follows: Section 2 presents framework of text mining technique. In section 3, we have discussed methods for summarization of text. Details of the text mining techniques are discussed in section 4. In section 5, we discuss classification techniques that uses for text mining. In section 6, we discussed research challenges and directions towards improvements of the techniques for text mining. In Section 7, we presented applications of text mining and Conclusions are given in section 7.

2. TEXT MINING FRAMEWORK

In spite of the truth that privacy of data and data security are frequently used as equivalent words, they share all the additional supportive category of relationship. Pretty lot as a home security framework secures the safety and integrity of a family unit, data security policy is put in place to ensure data privacy.

Text Mining is the process of extracting fascinating information or knowledge or patterns from the unorganized text that are from diverse sources. As the text is in unorganized form, it is quite difficult to deal with it. Finding interesting information from the natural language text is the prime purpose of text mining. The text mining process is shown in below figure 1:-

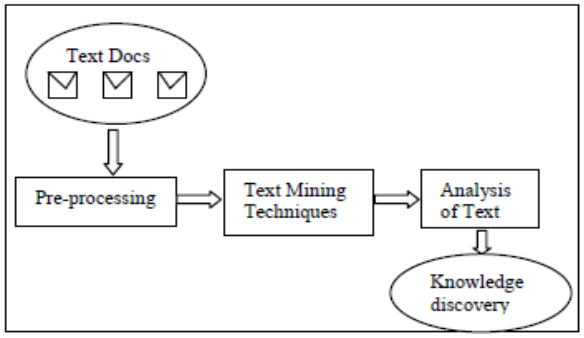

Figure1: Framework of Text Mining

In the below, we present state of the art approaches for the analysis of tasks preprocessing, classification, clustering and information extraction etc. The steps are as follows:

Step-I: Pre-processing Text: Mining from a pre-processed document is easy as compare to natural languages documents. So, pre-processing of documents that are from different sources is an important task during text mining process before applying any text mining technique. As Text documents can be represented as - a bag of words on which different text mining methods. To reduce the dimensionally of the texts words, appropriate methods such as filtering and stemming are used. Filtering techniques remove those words from the set of all words that do not give relevant information; stop word filtering is a conventional filtering method. After this step is applied, every word is represented by its root word.

Step II- Mining Technique: It is an important stage because in this step the selected algorithm is applied to text in order to process the document. The algorithm such as clustering, dimensionality reduction, knowledge representation and discovery, categorization, summarization, information extractions or visualizations could be used in general.

Step III - Analysis of Text: For knowledge discovery purpose, outputs which are coming from initial stage are analyzed here. For this purpose, various tools such as link discovery tool can be used. Here the unstructured text has been converted into some meaningful information from which one can make decisions

3. SUMMARIZATION OF TEXT

Due to the great amount of information, there is a need for producing summaries from number of documents. In this technique, the length of the text is reduced such that object and main points should not be lost. As summary can be given from a single document or group of documents and can replace the set of documents. A summary contains a significant position of information in the original document(s) and that is no longer than half of the initial documents.The process of summarization can be classified as:-

3.1 Pre-processing

In this step, the text is represented in a structured manner, by diminishing the dimensionally of the texts. This can be done by using filtering methods and stemming methods.

3.2 Categorization

In 1999 Yang, Y et.al.[20] approach, pre-defined groups are assigned to the text documents. The aim is to train the classifier on the base of known examples and then unknown examples are categorized automatically. A number of statistical classification methods can be applied to categorize the text.

3.2.1 Naive Bayes Classifier

Naive Bayes is a popular method for text categorization. With suitable preprocessing, it is competitive in this domain with more advanced methods including support vector machines. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. It depends upon the probabilistic relation between various categories. Naive Bayesian is simple and practical to implement as it implies that all the words of the records are independent to one another.

3.2.2 Nearest Neighbor Classifiers

In this kind of classifier, similarity of the unknown document is counted with all other document in the training set, if k similar records are considered then, it is called as k nearest neighbour classifier. The shortcoming of this method is the overhead of calculating similarly measure with respect to the entire document in the training set.

3.3 Clustering

Text Clustering is an unsupervised method in which no input out patterns is predefined. This method is based upon the idea of dividing the similar text into the same cluster. Individual cluster consists of number of records. The clustering is thought better if the contents of documents of intra cluster are more alike than the contents of inter-cluster documents. Clustering is a process used to group similar records but it differs from categorization in than documents are clusters on the fly rather of through the use of pre-defined topics .Clustering can be divided into two categories: hierarchical clustering and partitioned clustering.

The benefit of hierarchical clustering is that, it can use any form of similarity measure and the disadvantage is that once the clusters are formed, cannot be rebuilt, to improve performance, if needed.

3.3.1 K-means Algorithm:

This method can be applied to large data sets. The aim of the method is that k clusters are formed from the data set. Recursive updating of centroids is done in this method. Each cluster would have a reference point known as centroid which will be used in every round of iterations.

3.4 Information Extraction:

Natural language text documents contain information that cannot be used for mining. As documents are considered as a bag of words, they can be represented by vector model which then can be exercised as an input to the above-defined techniques such as classifications, clustering but this is not used for this method. In Information extraction, the

documents are first converted into the structured databases on which data mining techniques can be applied to extract knowledge or interesting patterns.

4. TEXT MINING TECHNIQUES

Various data mining techniques existed in literature are summarized as below:-

4.1 KNN based Machine Learning Approach

In 2014, V. Bijalwanet. al. [2] proposed KNN based Machine Learning Approach for Text and Document Mining. Information Retrieval (IR) is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web. The applications of IR are diverse; they include but not limited to extraction of information from large documents, searching in digital libraries, information filtering, spam filtering, object extraction from images, automatic summarization, document classification and clustering, and web searching.

The basic idea of KNN (K-nearest neighbour) method is to determine the category of a presented query based not only on the record that is nearest to it in the document space but in the categories of the k documents that are nearest to it. Having this in memory, the Vector method can be seen as an instance of the KNN approach, where k is 1. This work uses a vector-based, distance-weighted matching function, as did Yang, by measuring document's similarity like the Vector method. Below figure shows the sketch of the record indexing and retrieval method. From experiment, KNN shows the highest accuracy as compared to the Naive Bayes and Term-Graph. The drawback for KNN is that its time complexity is high but gives a better accuracy than others.

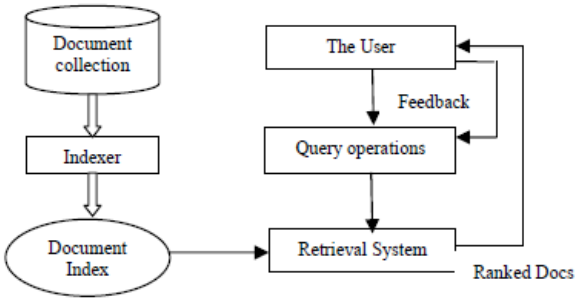

Figure2: Information Retrieval system

4.2 Hybrid feature selection based on enhanced genetic algorithm for text categorization

In 2016, Abdullah SaeedGharebet. al.[1] proposed Hybrid feature selection based on enhanced genetic algorithm for text categorization. This method uses a hybrid search technique that mixes the advantages of filter feature selection methods with an improved GA (EGA) in a wrapper procedure to handle the high dimensionality of the feature space and enhance categorization performance concurrently. First, we propose EGA by developing the crossover and mutation operators. The crossover operation is performed based on chromosome (feature subset) partitioning with session and document frequencies of chromosome entries (features), although the mutation is achieved based on the classifier execution of the original parents and feature importance. Thus, the crossover and mutation operations are performed based on useful information instead of using probability and random selection.

4.2.1 Enhanced genetic algorithm used

The GA is a good approach to explore the feature space and it can produce many alternative feature subsets through reproduction operations toward obtaining the best subset that includes the most important features. Several approaches have tried to improve the GA for FS, for instance by the use of subset size control in the fitness function and the use of the biological evolution concept with GA. However, the GA operators, i.e. crossover and mutation, are the key to achieving a diversity of the population and creating a new population for further runs of the algorithm. The adjustment of the probability rate of crossover and mutation is a difficult problem to solve and it is hard to coordinate these operations. The pseudo code of the suggested modification of the crossover and mutation operations given the paper [1].

4.3 Mind Map Generator Software Model with Text Mining Algorithm

In 2011, Robert Kudeliüet. al.[4] proposed Mind Map Generator Software Model with Text Mining Algorithm. A mind map is a picture used to represent words, ideas, tasks, or other items linked to and arranged around a central key word or idea. Mind maps are used to generate, visualize, structure, and classify ideas, and as an aid to studying and organizing information, solving problems, making decisions, and writing. Visual appearance of an example mind map is presented in [4].Everything relevant was extracted from the web page; therefore, this web data source was extracted satisfactory.

4.4 Rough Set Based Approach to Text Classification

In 2013, Libiao Zhang et. al.[5] proposed the Rough Set Based Approach to Text Classification. Textual document set has become an essential and rapidly growing knowledge source in the web. Text classification is one of the important technologies for information organization and management. Text classification has become more and more important and attracted the wide attention of researchers from different research areas. In this paper, many feature selection approaches, the implement algorithms and uses of text classification are introduced firstly.

However, because there is enough noise in the information extracted by current data-mining methods for text categorization, it leads to much uncertainty in the process of text classification which is created from both the information extraction and knowledge practice; therefore, more innovative techniques and methods are needed to improve the performance of text classification. It has been a critical step with the great challenge to further enhance the process of knowledge extraction and effectively utilization of the extracted knowledge. Rough Set decision-making approach is advised to use Rough Set decision techniques to more precisely classify the textual documents which are hard to separate by the traditional text classification methods.

4.5 A Fuzzy Based Approach

S. Goswamiet. al.[6] proposed a fuzzy based Approach To Text Mining And Document Clustering. In this paper, they have shown how to apply fuzzy logic in text mining in order to perform document clustering. Fuzzy logic is a mathematical logic paradigm in which truth can be biased i.e. it can have value between 0 and 1, that is totally false and

completely true. It is based on approximate reasoning instead of exact reasoning.

The FCM algorithm's iteration stops when the maximum change in the values of Fuzzy c-partition matrix is less than E, where E is a termination criterion with value between 0 and 1.

The main aim of this algorithm is to minimize the objective function given by:-

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \|^2, 1 \le m < \infty$$

Where,

m : denotes any real number greater than 1

u_{ij} : denotes membership degree of x_i in j^{th} cluster

x_i : is i^{th} dimension of the measured data

c_i : is the i^{th} dimension of the cluster centre.

For updating the cluster centre c_j , below formula is applied:-

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

The partition matrix values are updated using the formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$

The FCM algorithm's iteration stops when the maximum change in the values of Fuzzy c-partition matrix is less than E, where E is a termination criterion with value between 0 and 1.

4.6 A Text Clustering Method based on Huffman Encoding Algorithm:

In 2014, Maria Munteanet. al. [7] proposed a text Clustering method based on Huffman Encoding algorithm. In this, authors have suggested a new method for enhancing the clustering accuracy of text data. Our system encodes the string contents of a dataset using Huffman encoding algorithm, and declares these attributes as integer in the cluster evaluation phase. We gained a better clustering accuracy than the one obtained with traditional methods. This method is useful when the dataset to be clustered has only string attributes.Details of the Huffman encoding algorithm given in [7].

A. Concept based Mining Model

LincyLiptha R. et. al.[8] proposed text clustering method which is based on the mining model. It examines the terms based on the sentence, document and corpus level. The prototype consists of sentence-based concept investigation which determines the conceptual term frequency (ctf), document-based concept study which finds the term frequency (tf), corpus-based concept analysis which decides the document frequency (df) and concept-based similarity measure. The method of calculating ctf, tf, df, measures in a corpus is accomplished by the recommended algorithm which is called Concept-Based Analysis Algorithm. By doing so we cluster the web records in an effective way and the quality of the clusters achieved by this model significantly exceeds the traditional single term-base approaches Here, ctf represents conceptual term frequency, df represents document frequency.

5. CLASSIFICATION TECHNIQUES FOR TEXT MINING

Classification is the process of learning a set of rules from a set of examples in a training set. Text classification is a mining method that classifies each text to a certain category Classification can be further divided into three categories:

1. Machine learning based text classification
2. Ontology based text classification
3. Hybrid Approaches

5.1 Machine Learning based Text Classification

Machine Learning based Text Classification (MLTC) comprises of quantitative approaches to automate Natural Language Processing (NLP) that uses machine learning algorithms. Preferred supervised learning techniques for text classification are described in the subsequent text. Several techniques used for classification such as Rocchio algorithm based on feedback method, Instance based learning algorithm based new instances and already stored instances during training, Decision trees and Support vector machines, Artificial neural networks that contains supervised learning algorithms and genetic algorithms.

5.2 Ontology based text classification

Ontology can be the solution of the problems by introducing explicit specification of conceptualization based on concepts, descriptions, and the semantic relationships between the concepts Ontology represents semantics of information and is categorized as: (a) Domain Ontology consists of concepts and relationship of the concepts about a particular domain area, Basic components of ontology include: (a) classes, (b) attributes, (c) relations, (d) function terms, and (e) rules.

5.3 Hybrid Approaches

Several classification techniques [21] used for the text classification. Combination of different classification techniques used for the hybrid approach and that even provides efficient results in the text mining classification

6. RESEARCH DIRECTIONS AND CHALLENGES

In today's scenario, the amount of stored data has been enormously increasing, so to process that unstructured information several methods such as summarization, classification, clustering, knowledge extraction and visualization are available for the same which appears under the domain of text mining. Text mining also refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. In any Text mining or exploratory data analysis effort, visualization of textual data is an essential part of the problem. The scientific community has a great deal to contribute to many of these problems. So, we will mention some future research challenges in Text Mining as below:-

6.1 To made semantic analysis much more efficient and scalable for very large text is a research challenge

Intermediate forms with fluctuating degrees of complexity are suitable for various mining prospects. For a fine-grain domain-specific knowledge innovation task, it is necessary to conduct semantic analysis to conclude a sufficiently rich illustration to capture the relationship within the objects or concepts represented in the documents. However, semantic analysis approaches are computationally rich and often work in the order of a few words per second. It prevails a challenge to understand how semantic analysis can be done much more effective and scalable for vast text context.

6.2 Multilingual text refining is a major future challenge

Although data mining is largely language independent, text mining requires a significant language component. It is necessary to develop text refining algorithms that process multilingual text records and provide language-independent intermediate structures. While most text mining engines focus on processing English documents, mining from documents in other languages allows access to previously untapped information and offers a new host of opportunities.

6.3 It's interesting to explore how a user's knowledge can be used to initialize a knowledge structure

Domain knowledge, not provided for by any modern text mining tools could play a significant role in text mining. Definitely, domain knowledge can be applied as early as in the text cleaning stage. It is fascinating to investigate how one can take benefit of domain knowledge to enhance parsing efficiency and derive a further compressed intermediate form. Domain knowledge could also perform a part in knowledge distillation. In a classification or predictive modeling assignment, domain knowledge helps to enhance learning/mining efficiency as well as the quality of the learned model. It is also interesting to explore how a user's knowledge can be utilized to initialize a knowledge structure and make the discovered knowledge more interpretable.

6.4 Future Text mining tools should be more usable for inter-domain users

Modern text mining products and applications are still tools meant for qualified knowledge specialists. Future text mining tools, as part of the knowledge control systems should be readily applicable by technical users as well as administration executives. There have been some attempts in developing systems that express natural language queries and automatically implement the proper mining processes. Text mining tools could also look in the form of intelligent personal assistants. Under the agent paradigm, a personal miner would study a user's profile, conduct text mining operations automatically, and forward information without requiring an explicit request from the user.

6.5 Combining Rough set based machine learning to Text mining

Rough set based machine learning methods [5] can be applied to the domain of Text Mining. Rough set based knowledge representation is the newly evolved, growing and most efficient area in knowledge discovery as well as in artificial intelligence. A large number of high-quality papers on various aspects of rough sets and their applications have been published in recent years as a result of this attention. Since we know, most of the data of an organization is unstructured, and its dimension is more. So from this kind of data, text mining can be done by dimensionality reduction of that original data. The reduced data will also represent the same knowledge. So in this process of knowledge representation, we can convert the unstructured data into the meaningful information, so that one can take decisions from it. Text mining typically utilizes machine learning methods such as clustering, classification, association rules and predictive modeling. These methods uncover meaning and relationships in the underlying content.

7. APPLICATIONS

Text mining has several applications in the scientific community as well as in business domain. These applications are categorized as follows:-

7.1 Security application

Many text mining software packages are vend for security purposes, especially monitoring and investigation of online plain text references such as Internet news, blogs, etc.. It is also required in the study of encryption or decryption of the textual information.

7.2 In software Environment:

Text mining techniques and software is also being studied and developed by larger firms, including IBM and Microsoft, to further automate the mining and review processes, and by various firms working in the area of exploration and indexing in general as a way to improve their results. Within public sector much effort has been concentrated on creating software for tracking and monitoring terrorist activities.

7.3 In bio-medical field

Several text mining applications in the biomedical literature are existed. One online text mining application in the biomedical literature is PubGene that combines biomedical text mining with network visualization as an Internet service. TPX is a concept-assisted exploration and navigation tool for biomedical research analyses.

7.4 In marketing field

Text mining is rising to be used in marketing as well, more particularly in analytical customer relationship management.

7.5 In Academia

The subject of text mining is of interest to publishers who hold massive databases of information requiring indexing for retrieval. This is particularly true in scientific methods, in which highly precise information is often contained within written text.

7.6 In Sentiment analysis

Sentiment inquiry may involve analysis of movie reviews for estimating how favorable a review is for a movie. Such an investigation may need a labeled data set or labeling of the affectivity of words. The text has been used to detect emotions in the related area of affective computing. Text based approaches to affective computing have been used on multiple corpora such as students evaluations, children stories and news stories.

7.7 In Enterprise Business Intelligence:

Text Mining is an essential aspect of business intelligence that helps users and enterprises in analyzing stored text in a better way so as to make better decisions, improve customer satisfaction and gain competitive advantage. It is reliable than data mining as it gives deeper insight into the expanding

business area and extracts more productive data for business intelligence.

8. CONCLUSION

In this paper, we presented importance of text mining and survey of techniques used for text mining. Structured framework with categorization and classification techniques are also presented in the survey. We discuss the application of different text mining techniques for unstructured data sets reside in the form of text documents. We discuss the kind of techniques allows making a best search engine using database knowledge to work with filter, wrapper or even ontology.We have also described open areas and challenging problems, research directions in text mining. To conclude, there is a lack of widely-deployed techniques for automated analysis of un-structured text in the real world.

9. REFERENCES

[1] Abdullah SaeedGhareb, Azuraliza Abu Bakar, Abdul RazakHamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization", Expert Systems With Applications 49 (2016).

[2] VishwanathBijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual, "KNN based Machine Learning Approach for Text and Document Mining", International Journal of Database Theory and Application, Vol.7, No.1 (2014).

[3] DivyaNasa, "Text Mining Techniques- A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4,April (2012).

[4] Robert Kudeliu, MladenKonecki, MirkoMalekoviu, "Mind Map Generator Software Model with Text Mining Algorithm", 33 Int. Conf. on Information Technology Interfaces, June 27-30, (2011), Cavtat, Croatia.

[5] Libiao Zhang , Yuefeng Li, Chao Sun, WanvimolNadee, "Rough Set Based Approach to Text Classification", IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), (2013).

[6] SumitGoswami, Mayank Singh Shishodia, "A Fuzzy Based Approach To Text Mining And Document Clustering", arxiv.org/pdf/1306.4633.

[7] Maria Muntean, Lucia Căbulea, HonoriuVălean, "A New Text Clustering Method based on Huffman Encoding Algorithm", (2014) IEEE.

[8] LincyLiptha R., Raja K., G.TholkappiaArasu, "Enhancing Text Clustering Using Concept based Mining Model", IJECSE.

[9] A. Akilan, "Text Mining: Challenges and Future Directions", IEEE (ICECS '2015).

[10] S. S. Dhenakaran and S. Yasodha, "Semantic web mining: A critical review," International Journal of Computer Science and Information Technologies, 2011, vol. 2, no. 5, pp. 2258–2261.

[11] G. Stummea, A. Hotho, and B. Berendt, "Semantic web mining, State Of The Art And Future Directions A Knowledge And Data Engineering Group, University of Kassel, Institute of Information Systems, Humboldt University, Berlin, 2006.

[12] M. A. Aufaure, B. L. Grand, M. Soto, and N. Bennacer, "Metadataand ontology-based semantic web mining," in Web semantics & ontology, D. Taniar and J. W. Rahayu, Eds., 2006, pp. 259–296.

[13] G. Sampson, M. D. Lytras, G. Wagner, and P. Diaz, "Ontologies and the semantic web for e-learning," Educational Technology & Society, vol. 7, no. 4, pp. 26–28.

[14] Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[15] Navathe, Shamkant B., and ElmasriRamez, (2000), "Data Warehousing And Data Mining", in "Fundamentals of Database Systems", Pearson Education pvtInc, Singapore, 841-872.

[16] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg.

[17] Liu, F. & Lu, X. 2011. Survey on text clustering algorithm.In Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), China, 901-904, 2011.

[18] Luger, G. F. 2008. Artificial Intelligence: Structure and Strategies for Complex Problem Solving. 6th edn. Addison Wesley.

[19] Kano,Y., Baumgartner,W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L. &Tsujii, T. 2009. Data Mining: Concept and Techniques. Oxford Journal of Bioinformatics, 25(15), 1997-1998.

[20] Yang, Y. and Liu, X. (1999). "A Re-examination of Text Categorization Methods, in Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99), 1999, pp. 42-49.

[21] D.Q. Miao, Q.G. Duan, H.Y. Zhang, J. Na, Rough Set based Hybrid Algorithm for Text Classification. Expert Systems with Applications 36, pp. 8932-8937, 2012.

[22] Szymanski, J., Self-Orgaanizing Map Representation for Clustering Wikipedia Search Results. 2011.