International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.11-22
http://dx.doi.org/10.14257/ijdta.2017.10.1.02

A Survey on Text Mining Techniques and Methods: A Review Approach

Shivaprasad KM[1] and T Hanumantha Reddy[2]

Dept. of Computer Science and Engineering [1 2]
Rao Bahadur Y Mahabaleswarappa Engineering College[1,2]
Affiliated to VTU Belagavi
Ballari-583 104, Karnataka, India
shivakalmutt@gmail.com[1] and thrbly@gmail.com[2]

Abstract

Over last few decades, we have witnessed the enormous accumulation and usage of the data. Major issues faced by this data are mismatch and overload. The mismatch is the some useful or interesting data has been overlooked and overload is nothing but the gathered data is not one the user needed. To overcome this issue a technique of text mining has been developed. Text mining extracts the useful and interesting data from the large unstructured data; it helps to cope up with the issues. A complex task in text mining is the analysis and categorization of the extracted data. For the efficient and effective extraction and analysis of the patterns of data, various techniques and methods like categorization, clustering, summarization, stemming etc. have been recently developed. Some of the techniques and methods are discussed in this paper.

***Keywords**: We would like to encourage you to list your keywords in this section*

1. Introduction

In this modern digitalized world, there is rapid growth in the digital data, which is in semi structured, or unstructured data format, which has increased the trend in using this data and attracted towards processing the unstructured data into structured data. Today Text mining is an emergent area, which gained attention towards processing this kind of unstructured information into relevant information.

Text Mining is the process of discovering hidden useful and interesting pattern from the unstructured text documents. Text mining usually involves structuring the input text, deriving patterns within that structured data, and finally evaluation and interpretation of the output. Text mining produces the high quality (Relevant) data and eliminates the irrelevant data. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

The information retrieval from the unstructured data is very challenging task as it contains the massive information, which is needed to be processed. There are many methods and techniques for performing the text mining operation, which we have discussed in this paper

2. Techniques of Text Mining

2.1. Anomaly Detection

In this present world of huge data, it is very important to protect the data from the intruders and attackers. As the data is always transferred and shared it is usual that there is chances of it being exposed to attack. Techniques should be used to make the data less vulnerable. One such technique is Anomaly detection, which uses the concepts of data

ISSN: 2005-4270 IJDTA

mining to detect the surprising behavior within the data. It identifies the surprising activities within the data and compares it with normal activities.

Mainly anomaly detection is the process of finding the patterns within the data, which are not normal and usual. The data handled by these techniques are usually recorded data, which can be univariate and multivariate. These patterns are called anomalies or outliners. Detected anomalies may or may not be harmful. They should be properly studied and analyzed. The methodology used is first the normal profiles or patterns should be built which are taken as standard to detect the anomalies or outliners in the data. Immediately trigger is sent if any abnormal behavior is noticed. Relationship with the existing normal data should be made. Analysis of data can include various methods like classification, clustering, and other machine based learning techniques to know the critical and significant information. Thus eliminating the chances of intrusion [1].

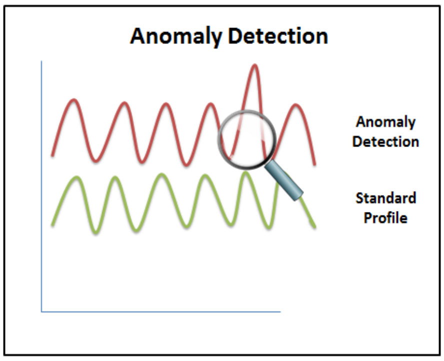

Figure 1. Anomaly Detection

2.2. Regression

In this training data is usually represented as pair of input and output matrices. The input matrix is identical to our feature matrix A and the output matrix B consists of flags indicating the category membership of the corresponding document in matrix A. Thus B has the same number of rows like A (namely m) and c columns where c represents the total number of categories defined. Regression methods are language dependent and they can be easily used with single and multiple category problems [2].

2.3. Classification

It is the method where it involves building up a model that can classify the objects so as to predict the future class. It can be done in two steps. In the first, the model is constructed based on the training set that defines the characteristics of classes and concepts the data in the training set. In the next step, the constructed model is used to predict the future classes. In other words the text classification can be called as dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes [3].

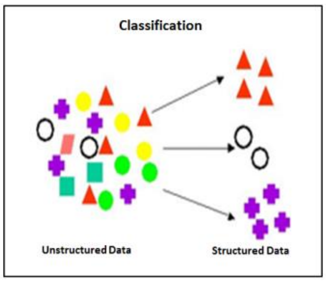

Figure 2. Classification

2.4. Clustering

Clustering of the document has been identified in different areas of text mining. This has become popular nowadays as its wide application in web mining, search engine, information retrieval. Clustering is a technique used to group similar documents. Document clustering is the process of organizing the documents into clusters or groups, documents within the clusters having the highest similarity compared to documents within other clusters. Representation of documents in the form of clusters necessarily loses some finer details but achieves the simplification. But it differs from categorization, in that documents are clustered on the fly instead of through the use of predefined topics. The Main goal of this clustering is to maximize the intra cluster similarity and minimize the inter cluster similarity. Advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results. But the challenge in clustering is to effectively identify the meaningful groups. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents [2, 4].

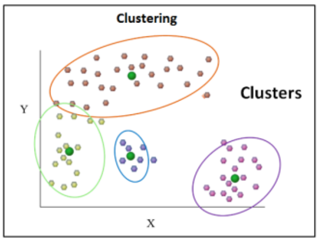

Figure 3. Clustering

2.5. Association Rule Mining

Association mining is a very important component of data mining. The relationships discovered by the data mining are expressed as association rules. Association rules are an important class of methods of finding patterns in data. The main goal is to find an interesting association or correlation relationships among a large set of data items. Since there is a massive collection of data which increases continuously, many companies are becoming interested in mining the association rules. In addition to the rule, the association mining also calculates some statistics about the rule. Four statistical measures are usually used to define the rule and these are the Confidence in the association, the Support for the association, the Lift value for the association and the Type of the association. Association mining has many application domains. Traditionally, the technique has been used to perform market basket analysis. Association rule mining has a wide range of applicability such market basket analysis, medical diagnosis, research, Website navigation analysis, homeland security, education, financial and business domain and so on [5-6].

2.6. Feature Selection and Extraction

The Feature Selection approach focuses on identifying relevant data, helps to understand and visualize the data, it also reduces the training and processing time of huge amounts of data as well as increase the accuracy for the subsequent data mining tasks. Feature selection is a long existing, novel method which aims to remove that irrelevant and noisy information by focusing and contour only the relevant and informative data for use in the text mining. Feature selection approach opens new research door for text mining. There exist two questions in feature selection approach first question is 'what are the features for machine learning which can represent the text in the effective way? And the second is: 'what is the best way to prune a large set of features down to a manageable set of most discriminating features?' For the first question we can say, it depends upon the processing power, language and corpora working with and most importantly the specific problem you are tackling .For the second question: We can try various approaches in order to prune the feature sets including: classifiers which classify and build the relevant information, Point wise Mutual Information to measure the association of an attribute to a class, finding the features those are strongly relevant and irrelevant and chi-square method can be used for evaluation of difference between the feature sets arised. In the broader way, there are two approaches for selecting the 'best' features, and they are filter and wrapper approaches. In filter method, selection of the subset is done without considering any algorithm which is usually done before processing. In wrapper method, evaluation of feature set is done by considering the algorithm [7].

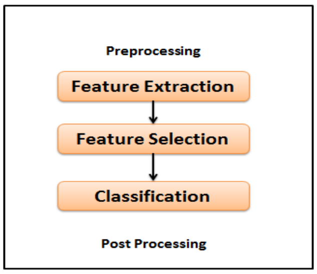

Figure 4. Feature Selection

2.7. Information Extraction

Analysis of the unstructured information is very much essential. This requires the information extraction. Much information is available in natural language rather than structural databases. This Information extraction identifies the key phrases, relationships within the text by the process of pattern matching. It transforms the corpus of textual documents into a structural form of databases. Once the information is extracted, it can be stored in the databases to be mined, queried summarizing in natural language, *etc.* First step in processing this extracted data is linguistic pre-processing. There are many linguistic techniques such as tokenization, part of speech tagging, lemmatization *etc.* in order to feed the information extraction system. IE is very useful for dealing with large volume of data [2, 8].

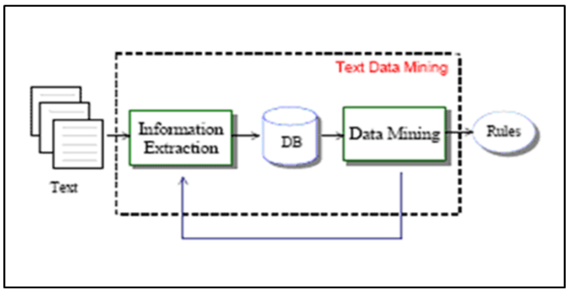

Figure 5. Information Extraction

2.8. Categorization

Usually the data in natural language form needs processing. Text Categorization does the assignment of those natural language documents to the predefined categories depending upon the content. Categorization engages identifying the main themes of a document by placing the document into a pre-defined set of topics. These sets of predefined categories are called as a controlled vocabulary. Categorization mainly relies on the vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, synonyms, and related terms. Categorization calculates words that emerge and, from the counts, identifies the main topics that the document covers. Categorization helps in grading the documents such that which documents have the most data on specific topics [2, 4].

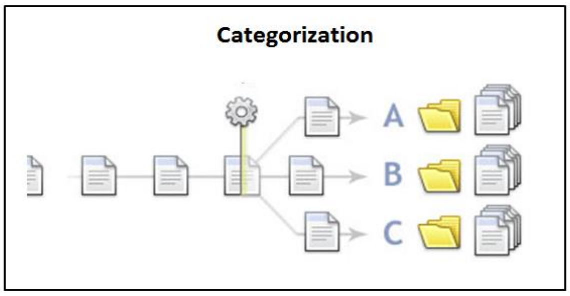

Figure 6. Categorization

2.9. Summarization

Summarization is a technique which is helpful usually in the very large documents. Providing the summary of the data makes the user to easily understand. It helps the user by immense processing and Figure it out whether the document meets the user needs and it is worthful to read further. The Main idea of summarization is to reduce the length, but retaining its main points and meaning. But the challenge is to analyze semantics and to interpret the meaning [2].

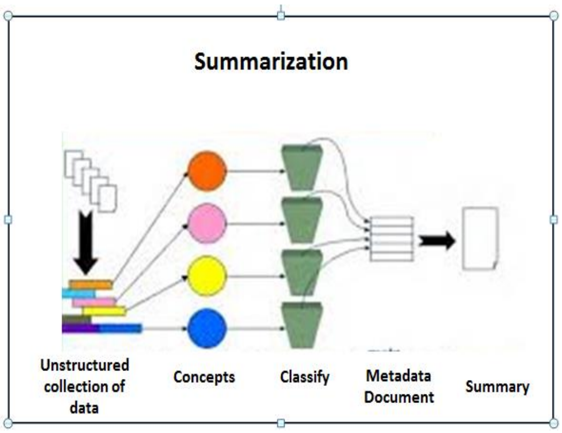

Figure 7. Summarization

2.10. Visualization

Rather than simple searching, Visual Text mining or information visualization makes the visual hierarchy or map of the large textual sources. Visual mining is the process of discovering the implicit but useful information from the large sets using the visual techniques. Popularly used tool in this technique is DocMiner allows the user to analyze the content visually. User can interact with the document by zooming, scaling and creating sub-maps. Visualization is more useful when user requires narrow down of the broader range of documents to explore the related topics [4].

2.11. Topic Tracking

The Topic tracking system helps the user in finding the information by the keywords and informs them of the related information. Based on the user profiles, user views it identifies the documents that are of interest to the user. There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly, businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments for illnesses and who wish to keep up on the latest advancements. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest [4, 9].

2.12. Concept Linkage

Concept linkage is another technique of text mining where it attempts to relate the documents by identifying the common shared idea or concept between the documents. This idea of the technique is very valuable in text mining, where it helps the user to browse the information rather than searching [4].

2.13. Question Answering

Question answering is another area of processing the natural language queries or question answering. It tries to find out how to answer the question in the best way. It uses the concept of information extraction to extract the entities and tries to categorize the questions for assigning them to the appropriate form (Who, What, When, Where, How). The Question Answering system takes in a natural language (NL) question from the user. This question is then passed to a Part-of-Speech (POS) tagger which parses the question and identifies POS of every word involved in the question. This tagged question is then used by the query generators which generate different types of queries, which can be passed to a search engine. These queries are then executed by a search engine in parallel. The search engine provides the documents which are likely to have the answers we are looking for. These documents are checked for this by the answer extractor. Snippet Extractor extracts snippets which contain the query phrases/words from the documents. These snippets are passed to the ranker which sorts them according to the ranking algorithm. Question Answering has various applications, helps user to find answers to common questions and some Frequently Asked Questions (FAQ) [9].

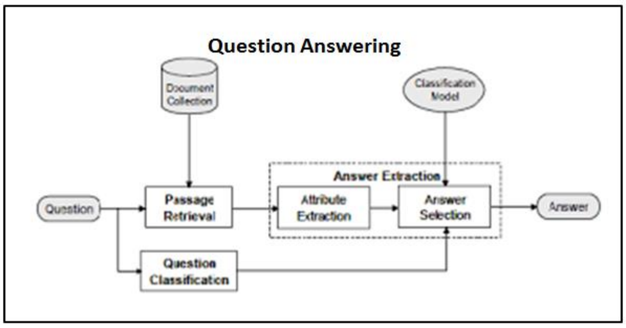

Figure 8. Question Answering

2.14. Stemming

Before the information can be retrieved, stemming process is applied to reduce the size of the document which can increase the effectiveness of information retrieval system. Words appearing in the documents or queries often contain the morphological variants.

These morphological variants have the similar semantic interpretations and can be considered as equivalent. Stemming is a process where these variant forms of words are reduced to a common form. It reduces inflected words to their stem, base or root form. There are a number of stemming Algorithms or stemmers, which attempt to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size. It results in smaller storage space and processing time. [10] There are stemming algorithms which differ in performance and accuracy.

The some of the different stemming algorithms are:

- Paice/Husk Stemming Algorithm
- Porter Stemming Algorithm
- Krovetz stemming Algorithm and
- Dawson Stemming Algorithm

All printed material, including text, illustrations, and charts, must be kept within the parameters of the 8 15/16-inch (53.75 picas) column length and 5 15/16-inch (36 picas) column width. Please do not write or print outside of the column parameters. Margins are 3.3cm on the left side, 3.65cm on the right, 2.03cm on the top, and 3.05cm on the bottom. Paper orientation in all pages should be in portrait style.

3. Methods of Text Mining

3.1. Term Based Method

Term based method is one in which document is analyzed based on the term. This term in the document has the semantic meaning. It enhances the efficient computational performance and mature theories for the term weighting. The Challenge faced in this method is, it suffers from the problems of polysemy and synonymy. Polysemy is a word having multiple meanings and synonymy is multiple words having the same meaning [11].

3.2. Phrase Based Method

Phrase method is one in which document is analyzed on the basis of the phrase. The Phrase has more semantics, less ambiguous and more discriminating than individual terms. Challenges faced in this method are when phrases have inferior statistical properties of terms, low frequency of occurrence, and when a large number of redundant and noisy phrases are present [11].

3.3. Concept Based Method

In this case of concept based method, the document is analyzed by the terms and concepts. Most of text mining techniques are based upon the statistical analysis of word or phrase. This statistical analysis captures the importance of the word or concept without the document. In the document two terms can have the same frequency, but only one term contributes the meaning appropriately than the other [11].

The concept based model includes three components, first analyses the semantic structure of the sentences. A Second component describes the semantic structures and the last component extract the concepts based upon the first two components. These concepts can be used to build feature vectors using the standard vector space model. Concept based model usually relies on the natural language processing. It can be effectively used to discriminate between meaningful terms which describe the sentence meaning and unimportant terms.

3.4. Pattern Taxonomy Method

In the pattern taxonomy method, the documents are analyzed based on patterns. Patterns can be structured into the taxonomy by using a relation. Patterns can be discovered by many techniques such as association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining [11].

There are two important stages in this method *i.e.*, extraction of the meaningful patterns from the text documents and how to make use of this discovered patterns for increasing the effectiveness. By splitting up the text documents into paragraphs and treating each paragraph as individual transaction, we can apply the method and generate the pattern taxonomies. This Pattern based method uses the two methods, pattern deploying and pattern evolving.

We need to interpret the discovered patterns. Feature space which consists of set of individual terms generated by the use of traditional document indexing techniques. Created pattern taxonomies and feature space can be used to represent the concept of documents by applying a data mining-based method like SPM. By deploying patterns into the feature space, PDM use of sequential patterns to keep the useful semantic information, but also improve the system efficiency by preventing the time-consuming pattern discovery approaches.

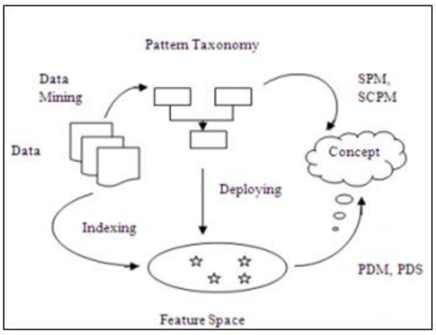

Figure 9. Pattern Deploying

The PTM has been significantly improved after the adoption of a pattern deploying method, which uses the strategy of mapping discovered patterns into a hypothesis space for solving the low-frequency problem to the specific long patterns. There is sometime negative documents contain some useful information to identify ambiguous patterns in the concept. A negative document nd is a document that the system sometime identified as a positive document. The offender of nd is a deployed pattern which obtains at least one component that appears in nd. When a negative document is detected, DPE starts to find offenders and implements pattern evolving at “Hypothesis Space” state. In contrast, IPE executes the same action at “Pattern” state. In addition, the structures of “Hypothesis Space” and “Pattern” are different, and thus an alternative definition and the algorithm for IPE are needed [12].

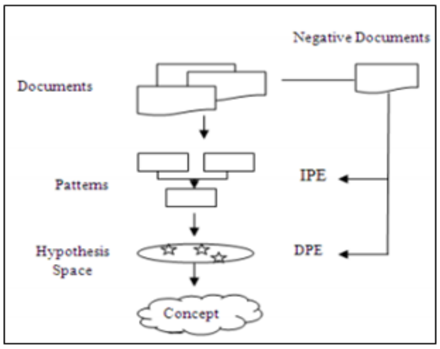

Figure 10. Pattern Evolving

Usage of discovered patterns is difficult and ineffective, as long patterns with low frequency may lack in support. But, all frequent patterns are not useful and may lead to misinterpretations ad ineffective performance. For the effective performance, effective pattern discovery technique has to be chosen to overcome the low frequency and misinterpretation problems.

4. Applications of Text Mining

Text mining is the interdisciplinary field with wide applications. Text mining is the user centric process with leverage analysis techniques and computing power to access the valuable information from the unstructured data. It eliminates the need to manually read the unstructured data. Today work on text mining is carried out on various domains [13].

Some of the applications of text mining:

- **Bioinformatics:** Bioinformatics literature is the target for text mining in this field. The Main goal is to allow the researchers with effective and useful knowledge discovery.
- **Business Intelligence**: Text mining is used by many organizations in the decision making. It makes analysts to directly arrive at the solutions by extracting and providing them only relevant data [14].
- **National Security:** Text mining is used as a surveillance tool. It provides the tools, methods, techniques, concepts to Figure ht against the terrorism, crime and helps in identifying future technical and operational challenges [15].
- **Knowledge Management:** As the information grows quickly, it is always challenging to manage this tremendous amount of data. Knowledge management software is provided based on text mining to provide the reliable solution.
- **Customer Care Service:** It is one of the traditional application of text mining. Many sources of information of customer are used such as reviews, surveys and provide a timely satisfiable response.
- **Fraud Detection through claims investigation:** Combines the effects of text analysis and structured data to prevent the frauds.
- **Content enrichment:** Helps in effectively managing the large volumes of information. Text mining techniques enrich the content, organize and summarize the available content that makes it suitable for a variety of purposes.
- **Spam filtering:** Today spam is a major issue for internet service providers and it is an entry point for the viruses for Text mining techniques can be implemented to improve the effectiveness of statistical based filtering methods**.**
- **Social media data analysis:** Today social media is one of the most prolific sources of unstructured data Text analytics can address both by analyzing large volumes of unstructured data, extracting opinions, emotions and sentiment and their relations with brands and products.

5. Challenges of Text Mining

- Major Challenges issue in text mining arises from natural language itself. Mining the unstructured data is often challenging because it is inconsistent and large.
- In the text documents one word may have multiple meanings as well as different words have the same meaning in statements resolving such kind of ambiguity in the document is one of the challenging issue of text mining (Ambiguity Problem).
- There is no uniform access over all sources. *e.g.* Web, email, databases, *etc.*, Preprocessing of data is required *i.e.*, necessary to reformat the text which is expensive and time consuming.
- Complex and Subtle relationship exists between the concepts in text mining.
- Learning techniques for processing text need annotated training. It is essential to build the text refining algorithms that process with multilingual text documents and produce language independent intermediate form.
- Another big challenge is how to make semantic analysis more efficient and scalable for every large corpus text document.
- Integration of domain knowledge in the text mining tool which plays an important role in the process of discovering the knowledge of documents is also an one more challenge of text mining.
- Another issue with text mining is cleaning the extracted data of online texts since reference address of the image link is difficult to remove.

6. Conclusion

Text Mining is the emergent, growing technology with its own inherent properties which can be used to solve the variety of problems today. It does not have any particularly certain techniques. Text Mining includes a wide variety of techniques and methods based on the problem considered and some of them discussed in this paper. These techniques and methods have undergone with various research studies.There are lots of challenges and opportunities found in this area. Text mining techniques have various applications and it is effective way of knowledge discovery without the particular domain knowledge.

Acknowledgments

I like to thank my guide Dr. T. Hanumantha Reddy sir for providing excellent guidance, encouragement and inspiration throughout the paperwork. Without his invaluable guidance, this work would never have been a successful one. In addition, I would like to thank my family & friends, who have been a source of encouragement and inspiration throughout the duration of the paper.

References

[1] S. Agrawal and J. Agrawal, “Survey on Anomaly Detection using Data Mining Techniques”, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, vol. 60, **(2015)**, pp. 708-713.

[2] G. R. Banu and V. K. Chitra, “A Survey of Text Mining Concepts”, International journal of innovations in engineering and technology, ISSN-2319-1058, vol. 5, no. 2, **(2015)**.

[3] Q. Zhao and S. S. Bhowmick, “Association Rule Mining: A Survey, Technical Report”, CAIS, Nanyang Technological University, Singapore, No. 2003116, **(2003)**.

[4] R. Patel and G. Sharma, “A survey on text mining techniques”, International Journal of Engineering and Computer Science ISSN:2319-7242, vol. 3, no. 5, **(2014)**, pp. 5621-5625.

[5] M. Tiwari, R. Singh and S. K. Singh, “Association–Rule Mining Techniques: A general survey and empirical comparative evaluation”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, no. 10, **(2012)**.

[6] I. Tudor, “Association Rule Mining as a Data Mining Technique”, Universitatea Petrol-Gaze din Ploieşti, vol. LX, no. 1, **(2008)**.

[7] K. Nirmala and M. Pushpa, “Feature based text Classification using Application Term Set”, International Journal of Computer Applications (0975 – 8887), vol. 52, no. 10, **(2012)**.

[8] H. Karanikas, C. Tjortjis and B. Theodoulidis, “An Approach to Text Mining using Information Extraction”.

[9] V. Gupta and G. S. Lehal, “A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, **(2009)**.

[10] N. S. Giridhar, K. V. Prema and N. V. S. Reddy, “A Prospective Study of Stemming Algorithms for Web Text Mining”, Ganpat University Journal of Engineering & Technology, vol. 1, no. 1, **(2011)**.

[11] S. V. Gaikwad, A. Chaugule and P. Patil, “Text Mining Methods and Techniques”, International Journal of Computer Applications (0975 – 8887), vol. 85, no. 17, **(2014)**.

[12] R. Bhaisare and V. Nayyar, “Analysis of Effective Pattern Discovery with Text Mining in Business Based Application”, International Journal of Research (IJR) ISSN 2348-6848, vol. 1, no. 11, **(2014)**.

[13] S. Jusoh and H. M. Alfawareh, “Techniques, Applications and Challenging Issue in Text Mining”, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814 www.IJCSI.org, vol. 9, no. 2, **(2012)**.

[14] http://www.expertsystem.com/10-text-mining-examples/

[15] http://store.elsevier.com/Application-of-Big-Data-for-National-Security/Babak-Akhgar/isbn-9780128019672/

[16] N. Biswas, “Text Mining and its business applications”, **(2013)**.