Text Categorization in Data Mining: A Review

Rimpy Wadhawan[1], Saurabh Mittal[2]
M.Tech. Research scholar, Galaxy Global Imperial Technical Campus, Dinarpur, Ambala
Associate professor, Galaxy Global Imperial Technical Campus, Dinarpur, Ambala

ABSTRACT

Content-based document management system has found outstanding status in the field of Computer and Information Systems Engineering and Computer Science. There are some causes for this popularity of content-based management system. The first one is that documents are available in digital form at a very huge scale. The second one is that the human beings have normal wish to access them in a flexible way. Text Categorization which is also known as Text Classification or Topic Spotting. The significance of Text Categorization (TC) is that the most popular web search engines like Google, Yahoo, Alta vista, Web Searches, Bing and others use Text Categorization (TC) to search data and metadata through the employment of web crawlers and returns the optimal results. Also "Search Engine Optimization" is a newly emerging area of research in Computer Science which needs novel and advanced research in Text Categorization (TC).

Keywords: *Text Categorization, k-means, Data Mining.*

1. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

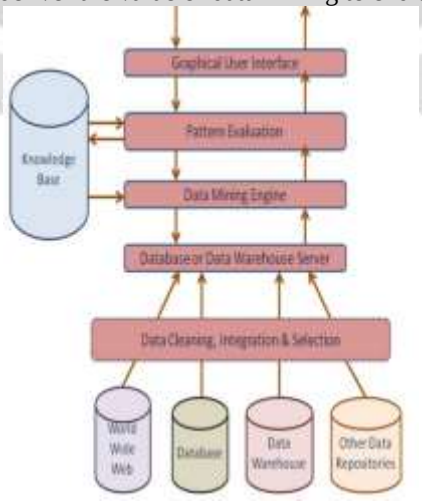

Fig 1- Data Mining Architecture

1.1 TEXT MINING

We have been given a predefined [9] set of natural language text then the method of labelling natural language texts with reference to thematic division is called Text Categorization (TC).There was an extensive work on Text Categorization in early 60s but this field was evolved gradually, and in early 90s it has gained prominent status and has become a major sub field of the Computer and Information Systems Engineering discipline. Obviously there is a role of increased power of software applications and the high availability of more powerful hardware in the emergence of Text Categorization (TC). There are now different applications of Text Categorization (TC) in many contexts. Some of the applications are Controlled Vocabulary Based Document Indexing, Document Filtering, Automated Meta Data Generation, Word Sense Disambiguation and Population of Hierarchical Catalogues of Web Resources. Generally speaking, Text Categorization (TC) is now being applied in multiple contexts covering any application requiring document organization, selective document dispatching and adaptive document dispatching. Text Categorization (TC) can be applied to the data which is in the form of natural language text. The natural language text is divided or categorized among subset of texts and labelled according to the theme which is the main idea or subject. Text Categorization is applied on online newspapers, online news channels, e-papers, web search engines because these web technologies incorporate search and retrieval of data in the form of text.

2. LITRATURE SURVEY

Mohammed G. H. et. Al. [1] presented an efficient rule-based method for categorizing free text documents. The contributions of this research are the formation of lexical syntactic patterns as basic classification features, a categorization framework that addresses the problem of classifying free text with minimal label description, and an efficient learning algorithm in terms of time complexity and F-measure. The framework of ROLEX-SP concentrates on capturing the correct classes of text as well as reducing classification errors. They performed experiments in order to evaluate the proposed method and compare our work with state-of-the-art methods in domain specific source of knowledge. The results indicate that ROLEX-SP outperforms other methods in terms of standard F-measure in medical domain because of the strong definition of MeSH description of medical categories.

Jiana Meng, Hongfei Lin, Yuhai Yu gives[2] proposes a two-stage feature selection algorithm. Firstly, they select features by the FCD feature selection method to reduce the feature numbers observably. Secondly, they apply LSI to construct a new conceptual vector space. The two-stage feature selection method conjugates the vector space model and the semantic feature space model. The proposed method not only reduces the number of dimensions drastically, but also overcomes the problems existing in the vector space model used for text representation. Through some applications involving spam database categorization, they find that our two-stage feature selection method outperforms other traditional feature selection methods.

Lam Hong Lee, Dino Isa, Wou Onn Choo , Wen Yeen Chue [3] gives High Relevance Keyword Extraction (HRKE) facility is introduced to Bayesian text classification to perform feature/keyword extraction during the classifying stage, without needing extensive pre-classification processes. In order to perform the task of keyword extraction, HRKE facility uses the posterior probability value of keywords within a specific category associated with text document. The experimental results show that HRKE facility is able to ensure promising classification performance for Bayesian classifier while dealing with different text classification domains of varying characteristics. This method guarantees an effective and efficient Bayesian text classifier which is able to handle different domains of varying characteristics, with high accuracy while maintaining the simplicity and low cost processes of the conventional Bayesian classification approach.

M. Maharasi et. Al. [4] proposed the compactness of the appearances of a word and the position of the first appearance of a word are used. Three types of compactness-based features and the position-of-the-first-appearance-based features are implemented to reflect different considerations. The distributional features are useful for text categorization, especially when they are combined with term frequency or combined together. The effect of the distributional features is obvious when the documents are long and when the writing style is informal.

Kostas Fragos , Christos Skourlas [5] proposed a method to improve performance in biomedical article classification. They use Naïve Bayes and Maximum Entropy classifiers to classify real world biomedical articles. They describe a technique based on chi-square measure to discard irrelevant information from the data and to identify the most relevant keywords to the classification task.

Shweta Taneja et. Al. [6] shown the major shortcomings affecting the traditional KNN algorithm and reviewed some improvements made to overcome them. Based on the analysis, they present our proposed KNN algorithm using dynamic selected, attribute weighted and distance weighted techniques. This proposed algorithm improves the accuracy of classification and reduces the execution time. It is a blend of classification and clustering techniques.

They have experimentally tested our algorithm in NetBeans IDE, using a standard UCI dataset-Iris. Experimental results have proved that our proposed algorithm performs better than conventional KNN algorithm.

Deqing Wang et. Al. [7] proposed a t-test feature selection approach based on term frequency. The student t-test is used to assess whether the averaged term frequencies of a term between two classes are statistically different from each other by calculating a ratio between the difference of two class means and the variability of the two classes. Then they compare our approach with the state-of-the-art methods on two common text corpora using three classifiers in terms of macro-F1 and micro-F1. Extensive experiments have indicated that our new approach offers comparable performance with v2, and ECE.

Anagha Kulkarni et. Al. [8] finds contexts of documents using pattern based clustering. Even though pattern based clustering is widely applied in gene expression or chromosome matching, it is not very common in text mining. The work proposed in this paper uses CSC to find contexts of clusters. The research reported in this paper suggests that CSC calculates closeness between patterns of documents. Experiments show that proposed algorithms are successful in finding context of unstructured text documents. The advantages of the proposed algorithms are: number of clusters are not given a priori, the clusters are resilient to noise and are formed only in one iteration. However, CSC calculation fails if weight in context vector is negative. Weights of context vectors are never negative in case of TF for text documents. Hence this condition is not explored in this work. However, in databases such as ionosphere, weights of context vectors could be negative. In such case, term-wise total of context vectors could be negative. CSC calculation needs modification in that case. WordNet is used to find context vector. Advantage of using low-dimensional context vector is that the matrix is not very sparse, the computations are time efficient and patterns are clear.

"Mahoshadha" [10] project focused on retrieving the most accurate answers to the any given query. According to the test results the goal has accomplished with 98% of accuracy and with a high efficiency of generating the response. "Mahoshadha" opens new paths in QA systems. Using both summarization and document clustering the response time can be improved. The described QA system does not rely on Named Entity Recognition, but focuses on only simple tagged corpus. With the use of SVM the authors introduce a technique in which the basic patterns can be identified. Using the pattern automatically identified the manual process can be heavily avoided, enabling to generate more accurate results, with avoidance of human error. The feature set will be provided to the SVM and then directly training the algorithm generated accurate results.

Goyal Shubham et. Al. [11] gives a methodology for the classification of sentiments was developed in this thesis for food price crisis in Indian market. Twitter API was used for streaming of tweets. The streamed tweets was filtered for relevant content and stored in a database. The several steps of pre-processing were applied on it and the tweets were removed from special characters, stop word, tokenized, etc. Stemming was done to all words in order to extract the root words. TF-IDF score based approach was utilized and the score was calculated for each tweets. Feature Selection was applied on it using Chi Square method and information gain. The extracted features form a term document matrix which is utilized in the classification algorithm. Two classification algorithms are compared as shown in previous chapter. The results are found to be satisfactory and when comparative analysis is done between them it is found that KNN outperforms Naïve Baye's Algorithm. Thus an automated system is designed for opinion mining related to food price crisis using Indian tweets.

3. CONCLUSION

With the dramatic rise in the use of the Internet there has been an explosion in the volume of online documents and electronic mail Text categorization the assignment of free text documents to one or more predefined categories based on their content is an important component in many information management tasks. We have studied different methods of text categorization in Data Mining.

VI. REFERENCE

[1] Mohammed G. H. Al Zamil , Aysu Betin Can, ROLEX-SP: Rules of lexical syntactic patterns for free text categorization, Knowledge-Based Systems, v.24 n.1, p.58-65, February, 2011 [doi>10.1016/j.knosys.2010.07.005].

[2] Jiana Meng, Hongfei Lin, Yuhai Yu, "A two-stage feature selection method for text categorization" in Computers and Mathematics with Applications 62 (2011), pp. no. 2793–2800.

[3] Lam Hong Lee , Dino Isa , Wou Onn Choo , Wen Yeen Chue, High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic, Expert Systems with Applications: An International Journal, v.39 n.1, p.1147-1155, January, 2012 [doi>10.1016/j.eswa.2011.07.116].

[4] M. Maharasi1, P. Jeyabharathi2, A. Sivasankari3, "Text Categorization Using First Appearance And Distribution Of Words", in Int. Journal of Engineering Research and Applications www.ijera.com Vol. 3, Issue 5, Sep-Oct 2013, pp.451-454.
[5] Kostas Fragos , Christos Skourlas, Toward Improving Classification of Real World Biomedical Articles, Proceedings of the 18th Panhellenic Conference on Informatics, October 02-04, 2014, Athens, Greece [doi>10.1145/2645791.2645848].
[6] Shweta Taneja, Charu Gupta, Kratika Goyal , Dharna Gureja, "An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering", in Fourth International Conference on Advanced Computing & Communication Technologies-2014,pp.-325-329.
[7] Deqing Wang, Hui Zhang, Rui Liu, Weifeng Lv , Datao Wan," t-Test feature selection approach based on term frequency for text categorization", in Pattern Recognition Letters 45 (2014) pp. no. 1–10.
[8] Anagha Kulkarni, Vrinda Tokekar, Parag Kulkarni, "Discovering Context of Labeled Text Documents using Context Similarity Coefficient" in Procedia Computer Science 49 (2015) 118 – 127.
[9] Ahmed Faraz"An Elaboration Of Text Categorization And Automatic Text Classification Through Mathematical And Graphical Modelling" in Computer Science & Engineering: An International Journal (CSEIJ), Vol.5, No.2/3, June 2015
[10] J. A. T. K. Jayakody, T. S. K. Gamlath, W. A. N. Lasantha , K. M. K. P. Premachandra, A. Nugaliyadde, Y. Mallawarachchi, ""Mahoshadha", The Sinhala Tagged Corpus based Question Answering System", in Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1, July-2016.
[11] Goyal Shubham, International Journal of Advance Research, Ideas and Innovations in Technology, ISSN: 2454-132X, Volume2, Issue5, pp. no. 1-9, 2016.