Rule Based Knowledge Extraction from Text Document using Probabilistic Algorithm for Text Mining

Ramesh Ranjan[1]
ram12ran@gmail.com[1]

Subhrata Roy[2]
subhrata1986@gmail.com[2]

Abstract— Due to heavy use of electronic device now a day most of the information stored in an electronic form. So the extraction of meaningful information from the large data set or a text document is challenge in text mining. Text mining is a process to extract the valid, novel, useful, fact data from the large text documents. Text mining essentially corresponds to information extraction and the extraction of facts from texts. There are several classification techniques. The probabilistic classifier has little bit older with indexing issue and probability calculation issue. In this paper, first applies the pre-processing on the text, and generates the rules. Second, calculate the probability according to these generated rule sets. The proposed method improves the accuracy, reduce false indexing and produce good result.

Keywords- Data mining, Text mining, Text Classification, K-mean, Categorization, probabilistic classifier

I. INTRODUCTION

Data Mining refers to use for extracting or mining knowledge from large amounts of data [1]. Data Mining is a process of discovering potential, useful, fact, novel, interesting and previously unknown pattern from large amount of data. With the use appropriate algorithm we can find out relevant information [1]. Sometime data mining is also referred as "Knowledge Discovery from Data (KDD)". There are many other terms similar to data mining such as knowledge extraction, data dredging, data archaeology. The information and knowledge gain can be use in market analysis, fraud detection, production control and scientific data analysis [1].

Text mining is one type of data mining technique. It use for extracting or mining knowledge from the text document. Text mining discovers the previously unknown information extracting it automatically from different source [2]. Text mining is similar to data mining. But the data mining dealing with structure data and text mining dealing with unstructured or semi structure data such as e-mail, text document and etc. The main objective of the text mining is to discover the previously unknown information. And the problem is that the result is not relevant to users need. In a text mining, the collection of documents from various different sources is easy but finding relevant information on demand is difficult.

Text mining process or text mining framework starts with the collection of document from different source. Text mining tools help to retrieve a document and perform preprocessing on it. Then document go to next stage it apply text mining techniques like classification, clustering, visualization, summarization, and information extraction. And the last step analyze the output data. For analyzing the output of text the users could navigate through in order to achieve the perspective.[3]

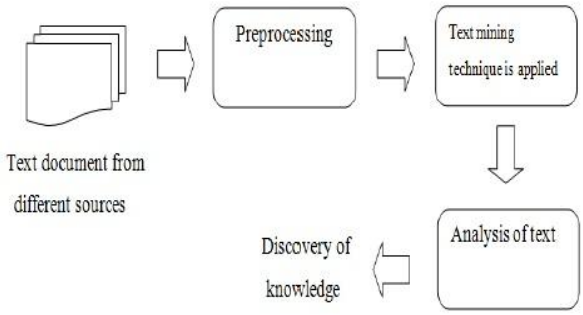

Figure 1: Text Mining Framework

II. TEXT MINING TECHNIQUES

Technology like information extraction, clustering, summarization, categorization and visualization are used in text mining frame work or process. Here in following section we discuss the text mining techniques [2].

A. Information Extraction

Information extraction is primary step for computer to analyze unstructured text and its relationship. This process is done by pattern matching is used to look for pre define sequence of text. IE is including identification, sentence segmentation. It is very useful for large text document. Many challenging in electronic information is in the form of natural language processing and IE solve this problem transform text document in to structure format.

B. Clustering

Clustering is unsupervised method. Clustering technique used to group similar documents but it differs from categorization, in this documents are clustered. This method is based on the concept of dividing similar text into same cluster. Each cluster contain a number of similar documents.

C. Summarization

Due to large amount of data we need to summarize the data from the number of document .which summarize the

Copyright © 2015 IJCSSCA |

data without change meaning of content, and the length of data. And produce summery from the group of document. Hence whole document set is replace by the summery. Summarization is helpful for the user read short summery of document instead of lengthy documents.

D. Visualization

In text mining visualization improve the simplicity to discover the information. Group of document or a single document text flag used to show document and color used. It provides faster batter and understandable information. It helps to discover or mine the pattern from collection of documents. Its use different color, relationship distance and etc.

E. Categorization

Categorization is similar to text classification [4] Categorization is a supervised technique because it is based on input output examples to classify. Text classifier is used to categorization of the text document in to pre define class. And pre define class is assign based on text document content. A typical text categorization process consists of preprocessing, indexing, dimensions reductions and classification. The goal of the categorization is to train classifier on the basis of known and unknown examples are categorized automatically. To categorize the text number of text classification techniques may use which are discussed further.

III. TEXT CLASSIFICATION TECHNIQUES

Text mining is a hot research area now a days. With rapid growing of it development Industry, business papers, eMAIL all data stored in electronic form so the large amount of data in and extracting a task relevant data from the large text document is difficult task. Here we are look some important text classification techniques which is basically use to categorize the text document into predefine class [4].

A. Nearest Neighbor Classifier.

KNN also called lazy learning or instance based learning. The KNN algorithm based on closest sample set. KNN is simple, valid and non parameter method. It is very easy to implement and need only two parameter. KNN is robust algorithm to deal with noisy data set. One of the major disadvantage is that its impossible to implement for large data set And cost become very high.

B. Support Vector Machine.

The SVM is popular high accurate machine learning method for text classification. SVM try to find an optimal hyperplane within the input space so as to correctly classify the binary (or multi-class) classification problem. SVM is less susceptible to over fitting than other learning method. It produces the best result for both test and training dataset. SVM is more complex to implement and cannot perform well in collection of text documents.

C. Association Based Classification.

Association based classification integrate association rule mining. Which generate class association rule and classification more accurate then decision tree and c4.5. Association based classifier is high classification accuracy and more flexible to Handel text data. A problem on classification is only based on support and confidence.

D. Centroid Based Classification

Cetroid based classification is mostly used. Its create centroid per class of the document. KNN is perform well but slow on the other hand centroid based classification is very fast because of similarity computation as the number of centroid need to be done. It is easy to implement and flexible for text data. Text collections are different number or sizes of document in class are unbalanced. So based on similarity we would like to classify. Based on document in class centroid based classifier select representative called centroid and it work k=1.

E. Decision Tree Induction

Decision tree is widely used inductive learning method. A popular decision tree classification algorithm is ID3, C4.5. A decision tree is like a flow chart or like a tree structure. Each branch represents the outcomes and node represents the test. Leaf node represents and holds a class label. Decision tree is simple and understandable dealing with noisy data. The algorithm may not guarantee for globally optimal decision tree because its greedy method perform locally.

F. Classification Using Neural Network

Neural network is important tools of text classification. It works well only when underlying assumption are satisfied. It is self adductive methods in that they can adjust data without explicit specification or distribution from for the underlying model. Application is fault detection, hand writing reorganization, speech reorganization medical diagnosis' and etc. its non linear model provide basis for established classification rule and performing statistical analysis. And more hidden nodes provide better classification.

IV. MOTIVATION

In many businesses circumstance, research is one of the main issues drives to the successful for obtaining the information from the outside of the organization. This information might be useful or not useful. The result will have different outcomes undertake from the different situation. Kotler (2006) defines research as "The Objective of exploratory research is to gather preliminary information that will help define problems and suggest hypotheses."

Copyright © 2015 IJCSSCA |

Text mining is an active research area. Most Text mining methods which produce a short text summary have not focused on the quantitative side. Future research can deal with this problem while also producing human readable texts And Extract Useful Information from Text Document. The current research is moving on different classification techniques for text mining such as KNN, SVM, Naïve Bayes, etc.

Probabilistic classifier used to extract information from text document. It calculates the probability values and according to value indexing the document to concern the group of cluster. This technique is now little bit older Problem in indexing, Lengthy process to calculate the probability of large text Document and competitively slow. There are many new methods and technique using which we can use and improve the performance of text mining accuracy and improve the extraction time.

V. PROBLEM STATEMENT

Most of the approaches are proposed towards a particular pattern mining and from smaller set of documents. Mining text from huge document set with different pattern is still an open challenge. Identify the text category from the large data set is pretty hard.

Applying text based classification techniques to categorize the text document in various categories. But, major issue is to extract the information in minimum accessing time and accurate relevant Information extraction from large text document.

The problem is discovering patterns and trends out of massive data are the great challenge. Any computer or laptop can accommodate huge amount of data due to advances in hardware storage devices. Accumulating information is easy but finding relevant information on demand can be difficult. In a previous work its hard to work on unstructured data set. Based on probabilistic classifier work on unstructured data is time consuming and difficult to find the probability of large text document. Its take more time to calculate probability. In a text mining Indexing is also Major problem for the text mining.

VI. PROPOSED WORK

Proposed system overcomes these Criteria. Reduce the Information Retrival accessing time. Increase the indexing Accuracy. And find the Accurate Probability from the Weight Factor in Text document. Here the Step We Extend the Existing Algorithm.

Step1: Convert unstructured data text file in to structure data text file
Step2: Create cluster of data text file by using k-mean.
Step3: Find the Centroid of the exiting nearby cluster and modify existing cluster.
Step4: Applying pre-processing
Step5: Rule Generation
Step6: Find the probability

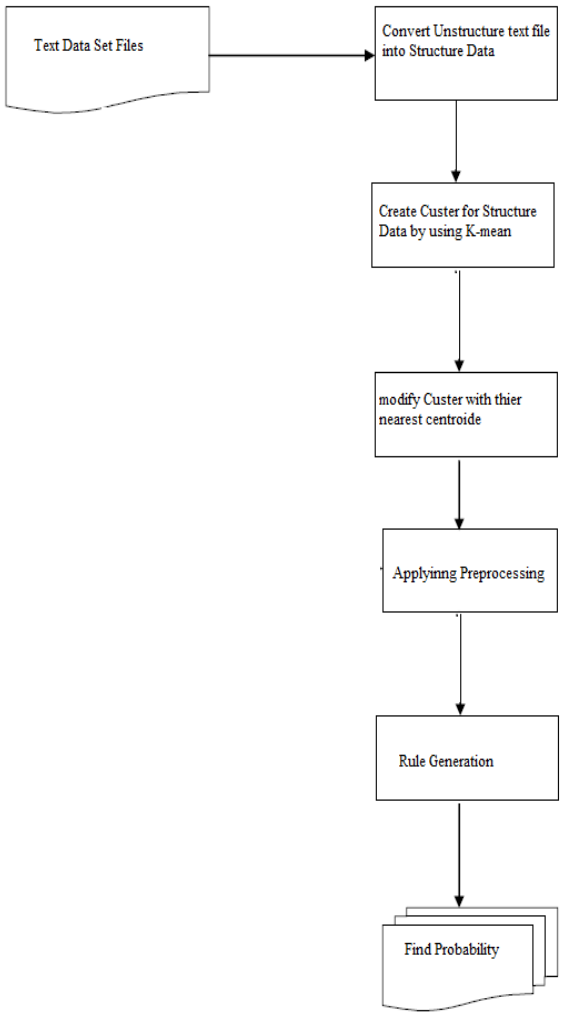

Figure 2: Flow Chart of Proposed System

VII. CONCLUSION

Text mining is a process of extracting knowledge from large text documents. And Extracting relevant information on demand can be difficult. There are several classifiers but With the use of efficient classification techniques, proper rule generation and modify the probability calculation we refine the Result. We extended algorithm which reduced the false indexing, Improve Accuracy, and easy to find the relevant information on demand. In a future work we proposed method based on k-mean algorithm, effective rule generation, Modify probability calculation to improve the overall result.

REFERENCES

1. Jiawei Han and Micheline Kamber “Data Mining Concepts And Techniques” ,Morgan kaufman publishers, San Francisco, Elsevier, 2011, pp. 285-351
2. M.Sukanyal, S.Biruntha2 "Techniques on Text Mining" International Conference on Advanced Communication Control and Computing Technologies, IEEE-2012
3. Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil "Text Mining Methods and

Copyright © 2015 IJCSSCA |

Techniques"International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014

4. Nidhi1, Vishal Gupta2"Recent Trends in Text Classification Techniques" International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, December 2011
5. S. Subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, IEEE-2013
6. M. Janaki Meena , K. R. Chandran "Naive Bayes Text Classification with Positive Features Selected by Statistical Method" ©2009 IEEE
7. vaishali Bhujade, N.J.Janwe "knowledge discovery in text mining techniques using association rule extraction" International Conference on Computational Intelligence and Communication Systems, IEEE-2011
8. Zhou Faguo, Zhang Fan "Research on Short Text Classification Algorithm Based on Statistics and Rules" 2010 Third International Symposium on Electronic Commerce and Security © 2010 IEEE
9. Shuzlina Abdul-Rahman, Sofianita Mutalib, Nur Amira Khanafi, Azliza Mohd Ali "Exploring Feature Selection and Support Vector Machine in Text Categorization" 16th International Conference on Computational Science and Engineering, IEEE-2013
10. Xianfei Zhang, Bicheng Li, Xianzhu Sun "A k-Nearest Neighbor Text Classification algorithm Based on Fuzzy Integral" Sixth International Conference on Natural Computation, IEEE-2010.

Copyright © 2015 IJCSSCA |