A Review on Knowledge Discovery using Text Classification Techniques in Text Mining

Chauhan Shrihari R
Computer Science and Engg. Dept.
Parul Institute of Technology, Limda, Vadodara

Amish Desai
Computer Science and Engg. Dept.
Parul Institute of Technology, Limda, Vadodara

ABSTRACT-

Data mining is process of identify the knowledge from large data set. Knowledge discovery from textual database is a process of extracting interested or non retrival pattern from unstructured text document. With rapid growing of information increasing trends in people to extract knowledge from large text document. A text mining frame work contain preprocess on text and techniques used to retrieve information like classification, clustering, summarization, information extraction, and visualization. . There are several text classification techniques are review in this review paper such as SVM, Naïve bayes, KNN, Association rule, and decision tree classifier. Which categorized the text data in to pre define class. In this review paper we study deferent techniques of text mining to extracting relevant information on demand. The goal of the paper is to review and understand different text classification techniques and finding the best one out for different prospective. From reviews I propose method with the use best classification method to improve the performance of result and improve indexing. And show the comparison of different classification techniques.

Keywords

Data mining, Text mining, Tex mining frame work, Text mining techniques, Text Classification SVM, Bayes, KNN.

1. INTRODUCTION

1.1 Data Mining

Data Mining refers to use for extracting or mining knowledge from large amounts of data.[1] Data Mining is a of process 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]. Data mining is also called "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]

1.2 Text Mining And Text Mining Frame Work.

Text mining is one type of data mining technique. The technique use for extracting or mining knowledge from

the text document. Text mining discover 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. Like eMAIL, text document and etc In a text mining main goal 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 document from various different sources. Collecting information is easy but fining relevant information on demand is difficult.

Text mining process or text mining frame work start 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] based on Following figuer1. Shows the Basic text mining frame work.

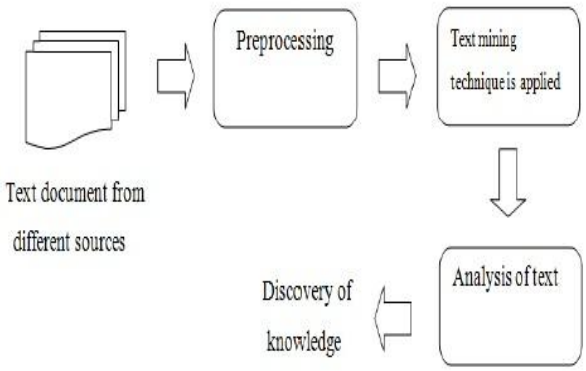

Fig 1. Text mining frame work

2. 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]

2.1 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 include identification, sentence segmentation. This techniques 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.

2.2 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.

2.3 Summarization

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

2.4 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. This method provide faster batter and understandable information. Which help to discover or mine the pattern from collection of document. Its use different color, relationship distance and etc.

2.5 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 consist of preprocessing, indexing, dimensions reductions and classification. The goal of the categorization is to train classifier on the basis of known and unknown example are categorized automatically. To categorize the text number of text classification techniques used which we will discuss in the following section (3)

3. 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]

3.1 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.

3.2 Bayesian Classifier.

It's a simple probabilistic classifier use to classify the text document. For text classification there are two different models of Naïve Bayes classifiers: Multi-Variate. Bernoulli Event Model and the Multinomial Event Model. Navie bayes is highly sensitive to feature selection. The navie bayes classifier is fast and easy to implement so its most popular and perform well. Its handle only low dimensions.

3.3 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. Its produce best result for both test and training data set. SVM is more complex to implement. And cannot perform well in collection of text documents.

3.4 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.

3.5 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. Its simple and efficient method. Its easy to implement and flexible for text data. Text collection are different number or size 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.

3.6 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 represent the outcomes and node represent the test. And a leaf node represent and hold a class lebel. Decision tree is simple and understandable dealing with noisy data. The algorithm can not guaranteed for globally optimal decision tree because its greedy method perform locally.

3.7 Classification Using Neural Network

Neural network is important tools of text classification. Its work well only when underlying assumption are satisfied. Its 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.

4. THE RELEVANCE OF KNOWLEDGE DISCOVERY USING TEXT CLASSIFICATION TECHNICES (LITRATURE SURVEY)

In [5] S. Subbaiah illustrated how to extract a knowledge from large text document. My initial study show that they proposed system which uses ODP taxonomy and domain ontology and dataset to cluster and identify the category of text document. Here they use probabilistic classifier (Naive Bayes classification) for text mining from text document. Proposed work is based on three step 1) Pre processing which pre process on input text document and removed stemmed, stop words, and split into paragraph and statement. 2) Rule generation here it generate positive and negative rule. 3) Probability calculation and generated positive and negative rule is used to calculate the probability value. According to probability value each term set or pattern are identified from text document. Based on probability value sort the positive and negative probability value and select the category from most top probability value. In this paper with the help of probabilistic classifier its generate good result but its have little false indexing. Here they used Reuters data set and each corpus data split in to ten category. They use 70% training data set and 30 % testing set. In future we generate a effective rule and change in probability calculation to improve the overall result of text mining.

In [6] M. Janaki Meena , K. R. Chandran, use naïve bayes classification techniques for specific positive features selected by statistical method. Paper proposed CHIR algorithm is supervised learning method for statistics which not define dependency of term but also define dependency of category is positive or negative. Algorithm begin with the pre processing which remove stop word and stemming word after stemming CHIR algorithm extract feature from training document. CHIR based algorithm is improve accuracy and most popular and simple classifier by proper identification of relevant information.

Vaishali Bhujade, N.J.Janwe knowledge discovery in text mining techniques using association rule extraction. In [7] automatically extract association rule from collection of textual document. The technique called EART. Its discover association rule among keyword labeling. ART system ignore the order of word its only focus on word. system based on TF-IDF And consist three phase 1) Pre processing. 2) Association rule mining algorithm for generate association rule based on weighted scheme.3) Visualization represent the result. The system is domain independent and flexible on different domain.

In [8] Zhou Faguo, Zhang Fan develop a method for short text based on rules and text classification. The algorithm mainly for feature extraction short text classification based on statistics and rule is proposed. and improve recall rate of text classification for short text.

In [9] Shuzlina Abdul-Rahman, Sofianita Mutalib, Nur Amira Khanafi Describe that finding the content from large text document is time consuming. Text categorization is process to assign a text in to pre define categories. The paper explore several feature selection that use to reduce dimension and feature space. The support vector machine adapt here and its fast and perform well. The accuracy is higher in feature selection, and ability to handel categorization problem for large data set.

In [2] M.Sukanya, S.Biruntha here Paper define Basic text mining definition , text ming frame work and its step how to text mining process works for extrxacting knowledge from text document. Text mining techniques like information extraction, clustering, visualization, and categorization. This is help in text mining. And visualization is used to provide better understandable information.

In [10] paper present the k nearest neighbor classification algorithm based on fuzzy integral it regards k-nearest sample set as k-evidance, which avoid independence demand of D- theory and improve performance. And experiment compare new method with improve knn classification algorithm. So the new method effective text classification algorithm with the combination of SVM can provide practical resolution for cosmic text classification.

Table I : Comparision Different Classification Techniques

Text classification techniques	***Precision (%)***	***Recall (%)***	***F1-measure(%)***
SVM	90.21	84.76	87.4
Bayes	79.83	80.74	80.05
KNN	78.33	80.26	79.28

5. CONCLUSION

After studying some papers related to knowledge discovery in text mining using text classification techniques. We analyze, that text mining techniques is very help full in the field of text mining, Day by day volume of electronic information is increase rapidly and extracting knowledge from these large volume data is difficult or say extracting relevant information on demand is very difficult due to large amount of data. So the main goal of text mining is to retrieve the relevant information in minimum accessing time, accurate data. For this purposed there are various approach, and techniques we will see in this survey paper. And with the efficient text classification techniques you can improve the text mining accuracy. Use any one of them, base on which techniques you are reviewing.

6. ACKNOWLEDMENT

With the cooperation of my guide, I am highly indebted to Asst. Prof. AMISH DESAI, for his valuable guidance and supervision regarding my topic as well as for providing necessary information regarding review paper.

7. 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 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 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

[7] 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

[8] 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

[9] 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