APPLICATION OF PATTERN MINING ALGORITHM FOR TEXT MINING

Rajvanshi Yadav, Siddhant Dongre, Nitin Sirsat, Advait save, Ganesh Gourshete
Department of Information Technology
Atharva College of Engineering, University of Mumbai, India.
{rjvnshi@reddiffmail.com}{dongresiddhant21,nitin.sirsat70,advaitsave, ganeshgourshete}@gmail.com

***Abstract --- There are* lots of data mining techniques have been present for mining useful patterns in text documents. However , it's an open research issue that how to update discovered patterns in unstructured text data. Since most existing text mining methods acquire term-based approaches, they all suffer from the problems of uncertainty and synonymy.**

Over the years, people have often thinks that pattern -based approaches should perform better than the term-based approach but many trials do not support this proposition. This paper presents a feature in new methods and effective pattern discovery technique which has the processes of pattern deploying and pattern evolving, to improve the productiveness of using and updating discovered patterns for finding relevant and important information.

***Keywords* — text mining, pattern mining, pattern evolving, pattern deploying, automatic categorization.**

I. INTRODUCTION

Due to the growth of knowledge created & gained in contemporary years, information detection and data processing have brought a good deal of observation with turning such information into meaning full data. Diverse applications, like market research and business management, will benefit by the engaging of the information and information extracted from an large amount of data. Information detection will be observed because the method of insignificant extraction of data from huge databases, information that's implicitly decorate within the information, formerly unknown and probably helpful for users. Data mining is so a important step inside the process of knowledge discovery in database. There are two main important stages are present in PTM, The First stage includes how to extract useful phrases from unstructured data. The second stage is then how to implement these discovered pattern to increase the effectiveness of a information discovery system. In PTM, we are going to split text document into a set of paragraph and use each paragraph as an separate transaction, which consist of set of words.

II.RELATED WORK

There are many types of text representations have been proposed in the past. In data processing techniques are used for text analysis by extracting coincidental terms as descriptive phrases from text collections. In data mining techniques have been used for text analysis by extracting co-occurring terms as descriptive phrases from document assemblies. The usefulness of the text mining systems using phrases as text representation showed no significant improvement.

Closed sequential patterns have been used for text mining in which proposed that the concept of closed patterns in text mining was useful and had the potential for improving the performance of text mining. PTM was besides developed to improve the efficiency by effectively exploitation closed patterns in text mining. a two-stage model used every term-based strategies and pattern based methods was introduced in to considerably improve the performance of data filtering

III. PROPOSED SYSYTEM

A. *Information Extraction*

The information extraction helps the computer to examine the unstructured text. We are using pattern matching, is the act of checking a perceived sequence of tokens for the presence of constituents of some patterns. When the size of the text is large. this technology is more useful than any other technology. It depends on the pattern that are extracted using the information extraction. The information extraction system selects the document that is applicable by analysing the incoming document to one or more query. There are two approaches for filtering namely term based approach and pattern based approach. conventional method is term based method which include positive document and negative document but this method focus on the positive document only. After pattern based method is also used but it will have a mediocre properties.

B. *Text Categorization(Training Dataset)*

Based on the content, the document can be allocated to already explain class. The major attribute of the text classification are the high importance of the feature space. The feature selection requires the indexing; tokenizing the text, feature space lessening. Knowledge engineering approach and machine learning approach are two main approaches in the text categorization. The knowledge engineering approach is defined by the user manually and the text is classified in given categories. Restricted access is one of the disadvantages of knowledge engineering approach which is also called as bottleneck in which the rules are defined manually. Whereas the machine learning approach is automatically assemble the text by knowledge a set of pre classified documents.

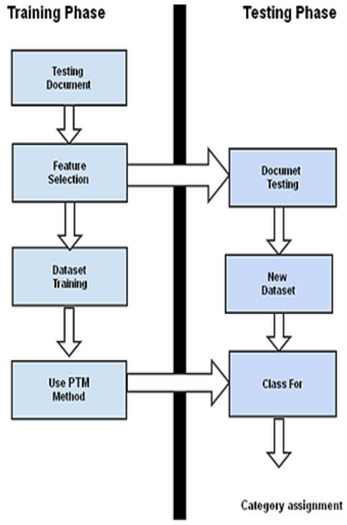

Fig.1. Text Categorization

Fig. 1 Text Categorization

1) *Keyword-Based Representation:*
Straightforwardness is the advantage of this approach. Whatever the words are extracted are kept in a feature space.
Similarity and dissimilarity are the disadvantage of this method. Choosing the limited number of features and modelling errors are another issue.

2) *Phrase-Based Representation:*
Phrases compose of more accurate content than single word. It can automatically find out the concealed semantic sequences of each classification in a documents; this can yield the classification correctness.

C. Pattern Taxonomy Model

Pattern methodology is the base of Pattern taxonomy model. This has two stages. In the first stage useful phrases are extracted from the text document. Close sequential pattern is obtained by using pattern taxonomy model and PTM only works on positive training document.

D. Pattern Deploying Method

Pattern deploying methods are planned for the use of knowledge discovered. All revealed patterns are not attractive because some noise patterns are also extracted from the training dataset. But negative document will also have some useful information to identify uncertain pattern in the concept . To enlarge the efficiency it is necessary for a system to utilize ambiguous pattern from the negative examples in order to decrease their influence.

E. Pattern evolving Method

In pattern evolving the patterns are updated. Complete conflict offenders and partial conflict offenders are the two main types of offenders. The complete offenders are get removed and partial offenders are get shuffled.

F. Automatic Categorization

Automatic text categorization is for all time a important application and research topic. It is useful for the huge amount of text document. It is the process of classifying a document in predefined category.

IV. METHODOLOGY

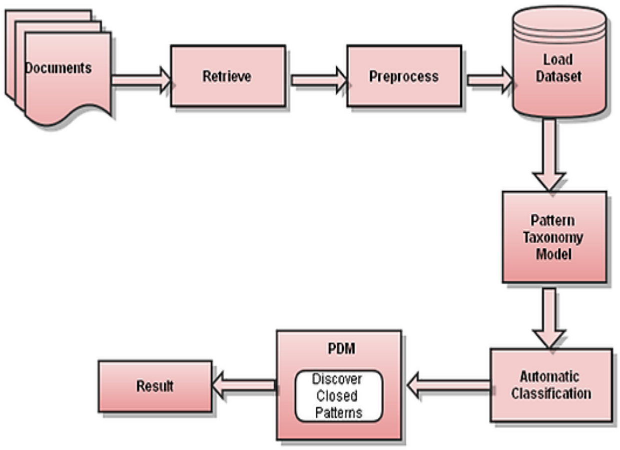

Fig.2: Overall Project Flow

1. *Loading document*

- To load a list of document.
- The user to recover one of the documents.
- This document is given to next process.

2. *Text Preprocessing*

- The recover document preprocessing is done in module.
- There are two types of process namely stop word removal and text stemming.

3. *Information Extraction*

- The information extraction helps the computer to examine the unstructured text.
- This technology is helpful when large volume of text is used.
- The key components of IE are set of patterns that are extracted.

4. *Text Categorization (Training Dataset)*

- The major characteristics of the text classification are the high dimensionality of the feature space.
- The feature selection involves the indexing, tokenizing the text, feature space lessening.

5. *Pattern taxonomy process*

- The documents are split into paragraphs and it considered as a document.
- The set of terms are mined from positive training dataset.

6. *Pattern deploying*

- The revealed patterns are summarized.
- Term support means weight of the term is evaluated.

7. *Pattern evolving*

- It is used to identify the noisy patterns in documents.
- Sometimes, system incorrectly recognized negative document as a positive.

8. *Automatic Categorization*

- Text categorization is the process of organizing a document in predefined category.

V. FUTURE SCOPE

Huge amount of information is produced everyday through economic, academic and social activities. To obtain the accurate pattern from large document pattern taxonomy model can be developed further to reduce the time to obtain the correct document. As the user is interested in a specific information that is important to him so ,he will not prefer the whole document.

VI. CONCLUSION

Many data mining techniques have been proposed in the last times. These methods contain connotation rule mining, frequent, sequential pattern mining, extreme and closed pattern mining. However, using these revealed information in the field of text mining is difficult and unproductive. The purpose is that some beneficial long patterns with high specificity lack in support. We argue that not all frequent short patterns are suitable. Hence, confusions of patterns derived from data mining techniques lead to the unsuccessful performance. In this examination effort, an effective pattern detection technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining.

ACKNOWLEDGMENT

This paper describes research done at Atharva College of Engineering in department of Information Technology. We express our gratitude to our project guide Mr. Ganesh Gourshete for guiding us. We are eager and glad to express gratitude to Head of Dept. Prof Jyoti Chinchole and all the Project coordinators. We would like to deeply express our sincere gratitude to our respected principal Prof. Dr. Shrikanth kallurkar and the management of Atharva College of Engineering.

REFERENCES

[1] *NingZhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.*

[2] *Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.*

[3] *S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" In iiWAS'05, pages 1013–1022, 2005.*

[4] *J.Wang and J. Han. BIDE, "Efficient Mining of Frequent Closed Sequences," Proceedings of the 2004 IEEE International Conference on Data Engineering (ICDE), pp. 79–90, 2004.*

[5] *Ning Zhong, Yuefeng Li, and Sheng-Tang Wu" Effective Pattern Discovery for Text Mining" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012*

[6] *M. J. Zaki. Spade: "An efficient algorithm for mining frequent sequences." Machine Learning, 42(1-2):31{60, 2001.*

[7] *"Knowledge Discovery in Text Mining Technique Using Association Rules Extraction" Bhujade .V,janwe CICN, 2011*

[8] *F. Sebastiani, "Machine Learning in Automated Text Categorization,"ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.*