International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Discover Effective Pattern for Text Mining

A. D. Khade [#1], D. S. Jadhav [#2], M. V. Vaghpatil [#3], A. B. Karche [#4], A. S. Zore [*5]

[#]*Student , Department Of Computer Engineering, University of Pune , MMIT, Lohgaon, Pune, Maharashtra, India.*

[*]*Lecturer, Department Of Computer Engineering, University of Pune, MMIT, Lohgaon, Pune, Maharashtra, India.*

Abstract:- **Multiple data mining techniques have been discovered for finding useful patterns in documents like text document . However, how to use effective and bring to up to date discovered patterns is still an open research task, especially in the domain of text mining. Text mining is the finding of very interesting knowledge (or features) in the text documents. It is a challenging task to find appropriate knowledge (or features) in text documents to help users to find what they exactly want. This paper represent efficient mining algorithm to find particular patterns within a reasonable and acceptable time frame.**

Keywords:- **text mining; text categorisation; pattern mining; pattern evolving.**

I. INTRODUCTION

In past few years, a number of data mining techniques have been presented to find out the different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. In this paper, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

II. EFFECTIVE PATTERN: THE CONCEPT

Text mining is the discovery of useful knowledge in text documents. It is a very difficult task to find accurate knowledge in text documents to help users requirement. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, The merits of term-based methods include efficient computational performance as well as some theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of IR. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include:
1) phrases have inferior statistical properties to terms,
2) they have low frequency of occurrence, and
3) there are large numbers of redundant and noisy phrases among them .

In the presence of these set backs, sequential patterns used in data mining community have turned out to be a promising alternative to phrases because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches have been proposed, which adopted the concept of closed sequential patterns, and pruned non closed patterns.

III. PROPOSED WORK

An effective pattern discovery technique is discovered. Evaluates specificities of patterns and then evaluates term-weights according to the distribution of terms in the discovered patterns Solves Misinterpretation Problem. Considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries to reduce their influence for the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evolution. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.

In General there are two phases:

Training and Testing:-

Training: In training phase the d-patterns in positive documents (D) based on a min sup are found, and evaluates term supports by deploying patterns to terms.

Testing: In Testing Phase to revise term supports using noise negative documents in D based on an experimental coefficient. The incoming documents then can be sorted based on these weights.

1) *Datasets:-*

Dataset is collection of the data which present in tabular form i.e. we can represent the data in row & column

wise format. Here we use special type of dataset in our system known as RCV-1(Reuters Corpus Value 1).

IV.SYSTEM ARCHITECTURE

The proposed network architecture is shown in Figure 1. This architecture shows the stepwise solution of our project. The basic step is to load documents in our database. The next step is to remove stop word and text steaming. we removed this stop word and text steaming with the help of NLP(natural language process).

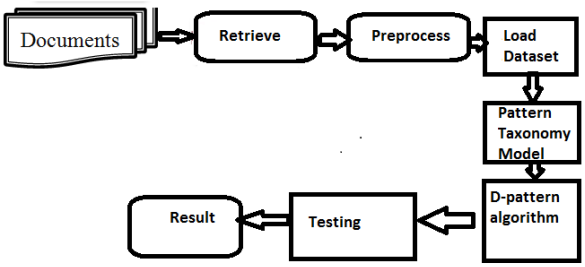

Fig 1: System Architecture

There are 5 sub modules of proposed system.
1) Loading documents
2) Text Preprocessing
3) Pattern taxonomy process
4) Pattern deploying
5) Pattern Testing

1) Loading documents
In this module, to load the list of all documents. The user to retrieve one of the documents. This document is given to next process. That process is preprocessing

2) Text Preprocessing
The retrieved document preprocessing is done in module.
There are two types of process is done.
a) stop words removal
b) text stemming
stop words are words which are filtered out prior to, or after, processing of natural language data. stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word form.

3) Pattern taxonomy process:-
In this module, the documents are split into paragraphs. Each paragraph is considered to beach document. The terms , which can be extracted from set of positive documents.

Where d=document;

m= set of paragraph;

s=keyword;

Paragraph	Terms
dm1	s1, s2
dm2	s3, s4 , s6
dm3	s3 , s4, s5, s6
dm4	s3, s4, s5, s6
dm5	s1 , s2, s6, s7
dm6	s1 , s2, s6 , s7

TABLE 1 :- A Set of Paragraph

Frequent Pattern	Covering sets
{ s3, s4, s6}	{dm2,dm3,dm4}
{ s3 , s4 }	{dm2,dm3,dm4}
{ s3 , s6}	{dm2,dm3,dm4}
{ s4 , s6}	{dm2,dm3,dm4}
{ s3 }	{dm2,dm3,dm4}
{ s4 }	{dm2,dm3,dm4}
{ s1 , s2 }	{dm1,dm5,dm6}
{ s1 }	{dm1,dm5,dm6}
{ s2 }	{dm1,dm5,dm6}
{ s6}	{dm2,dm3, dm4,dm5,dm6}

TABLE 2 :- Frequent Pattern and Covering set

4) Pattern deploying
The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated.

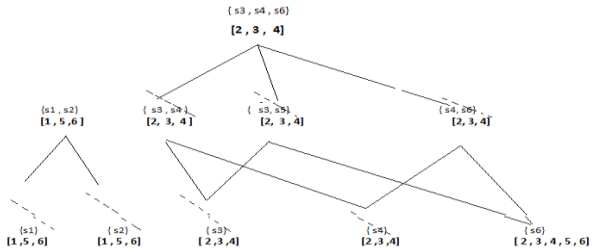

Fig.3 : Pattern Discovery

5) Pattern Testing

In this module used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. If partial conflict offender contains in positive documents, the reshuffle process is applied.

V. TECHNICAL OVERVIEW

A. Advantages

1. It improves the effectiveness of using and updating discovered patterns for finding relevant and interesting information.
2. Able to produce the text mining on polysemy and synonymy effectively.
3. Effective pattern Discovery Technique.
4. The proposed approach is used to improve the accuracy of evaluating term weights.
5. Because, the discovered patterns are more specific than whole documents.
6. To avoiding the issues of phrase-based approach to using the pattern-based approach.
7. Pattern mining techniques can be used to find various text patterns.

B. Disadvantages

1.Time consuming.

VI. EXPECTED RESULTS

To focus on the development of knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Technology can help in fast finding of text. There is efficient use of text mining. Finding searching pattern with there location in effectively. Processing and Multilingual Aspects are present in system.

VII. CONCLUSION

Thus mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining.

ACKNOWLEDGMENT

We would like to sincerely thank Mr. Amit Zore, our mentor (Lecturer, MMIT, Lohgaon), for his support and encouragement.

REFERENCES

[1] J. Han and K.C.-C. Chang, “Data Mining for Web Intelligence,” ,2002.
[2] A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003 http://www.textmining.org/home/
[3] Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization,” Information Retrieval.
[4] “Interpretations of Association Rules by Granular Computing,” By Y. Li and N. Zhong.
[5] Data set RCV1 . http://www.RCV1dataset.com/home/
[6] Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections by H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo.
[7] K. Aas and L. Eikvil, “Text Categorisation: A Survey,” Technical Report Raport NR 941, Norwegian Computing Center, 1999
[8] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.