International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 09 | Sep-2016 www.irjet.net p-ISSN: 2395-0072

Refined Pattern Discovery Model for Text Mining

RaviKiran.M[1], RamaDevi.P[2]

[1]*Student of M.Tech, Dept. Of Information Technology, VRSEC, Vijayawada, A.P, India*
[2]*Assistant Professor, Dept. Of Information Technology, VRSEC, Vijayawada, A.P, India*

Abstract – *Most of the data mining methods have been proposed for mining important patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue. Since most of text mining methods used term-based approaches, they all suffer from the problems of polysemy and synonymy. This paper presents an innovative and refined pattern discovery method which includes the processes of pattern deploying and pattern evolving, to refine the effectiveness of using and updating discovered patterns for finding relevant and interesting information.*

Key Words: Text mining, Polysemy, Synonymy, Pattern discovery, Pattern deploying, Pattern evolving.

1. INTRODUCTION

By the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with need for turning such data into useful information and knowledge. Knowledge discovery can be viewed as the process of important extraction of information from large databases, information that is completely presented in the data, formerly unknown and potentially useful for users. Data mining is therefore a fundamental step in the process of knowledge discovery in databases.

In the past years, a notable number of data mining techniques have been presented to perform various tasks for knowledge. These techniques include association rule mining, frequent item set mining, sequential pattern mining, and closed pattern mining. Most of these techniques are suggested for the purpose of developing efficient mining algorithms to find accurate patterns within an equitable and suitable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still a complicated issue. In this paper, we focus on the development of a knowledge discovery technique to effectively use and update the discovered patterns and apply it to the field of text mining.

Text mining is the finding interesting knowledge in text documents. It is a complicated issue to find accurate knowledge (or features) in text documents to help in users search for to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this issue, such as Rocchio and probabilistic models [4], rough set models [23], BM25 and support vector machine (SVM) [34] based filtering models. The use of term-based methods include efficient performance as well as sophisticated theories for term weighting, which have developed over the last couple of decades from the IR communities. However, term-based methods suffer from the problems of polysemy and synonymy, where polysemy implies a word has different implications, and synonymy is different words having the same implication. The semantic importance of numerous found terms is unverifiable for noting what clients need.

There are two fundamental problems regarding the viability of pattern-based approaches: low frequency and misinterpretation problems. Given a specified topic, a highly frequent pattern (commonly a short pattern with large support) is usually a general pattern, or a specific pattern of low frequency. In the event that we diminish the minimum support, a more number of noisy patterns would be discovered. Misinterpretation implies the measures utilized as a part of pattern mining (e.g., "support" and "confidence") turn out to be not relevant in using discovered patterns to answer what users want. The crucial problem hence is how to utilize the discovered patterns to accurately evaluate the weights of useful patterns in text documents. Many terms with bigger weights (e.g., the term frequency and inverse document frequency (tf*idf)) are general terms because they can be frequently used in both suitable and unsuitable information.

For instance, term "LIB" may have bigger weight than "JDK" in a sure of information accumulation; yet we trust that term "JDK" is more particular than term "LIB" for depicting "Java Programming Language"; and term "LIB" is more broad than term "JDK" since term "LIB" is additionally every now and again utilized as a part of C and C++. In this way, it is not satisfactory for assessing the weights of the terms in light of their dispersions in archives for a given subject, despite the fact that this assessing technique has been often utilized as a part of creating IR models.

Therefore, it is not acceptable for evaluating the weights of the terms-based on their conveyances in documents for a given topic, although this method has been frequently used in developing IR models. Hence discovered specifities of patterns are calculated and then evaluates term weights according to the sharing of terms in the discovered patterns rather than the sharing in documents for rectifies the misinterpretation problem. It also considers the impact of patterns from the negative training examples to find vague

(noisy) patterns and attempt to lessen their impact for the low-frequency problem. The process of updating vague patterns can be referred as pattern evolution. The process carried out in the current scenario is started with pre-processing technique that splits into terms, terms are stemmed, and stop words are removed. Terms are analyzed to discriminate the meaningful and non-meaningful terms. Positive and negative documents both undergo pattern taxonomy model technique for discovering closed patterns, frequent patterns and closed sequential patterns. Discovered patterns deployed summarized as d-patterns in order to find the term support based upon their occurrence in patterns rather than their occurrence in documents. With searching process done in frequent patterns, closed patterns, and closed sequential patterns, required information is discovered in form of terms. However time taken for extraction for terms will be calculated based on time complexity of each above mentioned technique.

2. LITERATURE SURVEY

Text mining is the technique which is helpful for the users to find the required information from a large amount of text data. It is therefore very important that the information retrieved for the users should be relevant and efficient. Term based methods were used earlier to overcome these issues. But the term based approach faces the problem of polysemy and synonymy. Polysemy implies a word has different implications, and synonymy is different words having the same implication. To overcome the drawbacks of term based method phrase based methods were developed. But phrase based approach also had some drawbacks like its it has inferior statistical properties to terms, frequency of occurrence of the phrases is low as compared to the keywords, it contains large number of noisy phrases and redundancy is also an issue [2][3].

The existing system uses the term based methods for extracting the text from the documents. These methods provide the text representations and pattern evolution techniques. Example of text representation can be hierarchical clustering which is generally used to determine the relations between the keywords [6][7]. Pattern evolution is useful in improving the performance of term based approach. The limitation of the term based approach is that a term with higher values could be sometimes meaningless in some d- patterns.

The proposed system mainly focuses on the specificities of the patterns. It then evaluates the weight of each term according to its distribution in the discovered patterns rather than the whole document. This weighting scheme avoids the problem of misinterpretation. The system take into consideration about the influence of patterns occurred from the negative training examples. This consideration is useful in determining the ambiguous or noisy patterns and it tries to reduce their overall influence. This also reduces the problem of low frequency. Two main process, pattern evolution and pattern deployment, is included in this system. Pattern evolution refers to the process of updating ambiguous patterns. These methods improve the accuracy of evaluating the weights of each term because discovered patterns are proved to be more specific than the whole document set. In general there are two phases training and testing. In training phase, the algorithm of pattern taxonomy model is invoked (PTM)[4][5]. This algorithm is used to find the d patterns in positive documents. In testing phase, the weights of all the incoming documents are evaluated. These documents are sorted and sorting is based on their weights [1].the accuracy of calculating the term weights can be improved with the proposed approach. Because the patterns discovered will be more specific instead of the whole document. To overcome the issues faced by phrase based approach, the pattern based approach is preferred and the pattern mining techniques are very useful in finding the text patterns.

3. PROBLEM STATEMENT

To discover patterns and then compute specificities of patterns for evaluating term weights as per their distribution in the discovered patterns form large amount of text document.

To improve the effectiveness by effectively using closed patterns in text mining. An effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation issues for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the found patterns in text documents. The advantage of the concept-based model is that it can successfully separate between non essential terms and meaningful terms which describe a sentence meaning. In this framework we are giving high need for long arrangement in the evaluated pattern. In this system we are giving term weight based on occurrence of term in long pattern (sequence).The proposed model outperforms not only other pure data mining based strategies and the concept based model, but also term based state-of-the-art models. The Proposed system will be the output of the project, it will include the discovering the patterns from the large amount of data and search for interesting patterns that user want.

4. PROPOSED SYSTEM

The proposed technique uses mainly two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. Our Proposed algorithm utilizes the semantic relationship between words to create concepts. Associating a meaningful label to each final cluster is more essential. Then, the high dimensionality of text documents should be reduced. Only authorized user retrieve the particular dataset for pre-processing and apply three

methods. When complete the process with three methods it will provide data with term set and their experimental coefficient to user as per their selected dataset. The step by step is showing in below flow diagram.

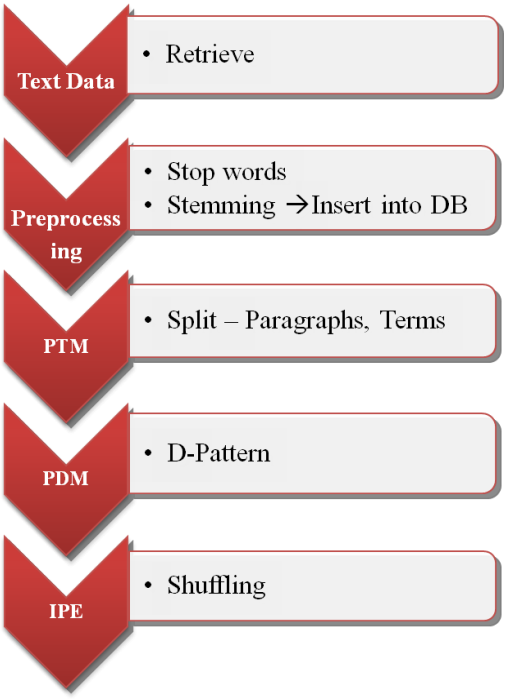

Figure 1: Flow Diagram of Refined Pattern Discovery

5. SYSTEM ARCHITECTURE

The proposed architecture is shown in Figure 2. 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 stemming.

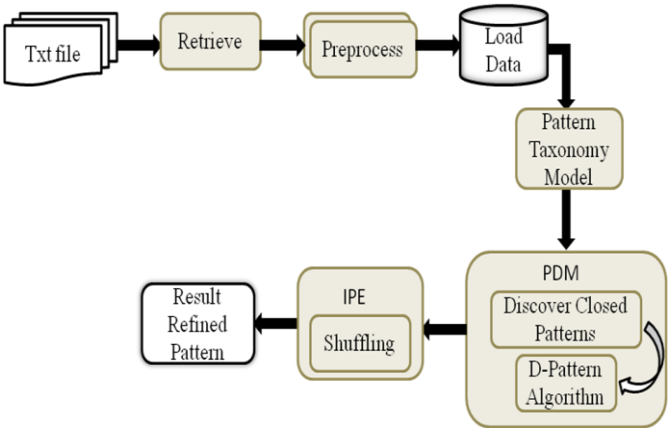

Figure 2: System Architecture Diagram

There are 5 sub modules of proposed system.
1) Loading Documents
2) Text Preprocessing
3) Pattern Taxonomy Model
4) Pattern Deploying
5) Pattern Evolving

5.1 Loading Documents

In this module, 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.

Figure 3: Loading Dataset

5.2 Text Preprocessing

The retrieved document preprocessing is done in module. There are two types of process is done.

a) Stop words removal

b) 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 written words form.

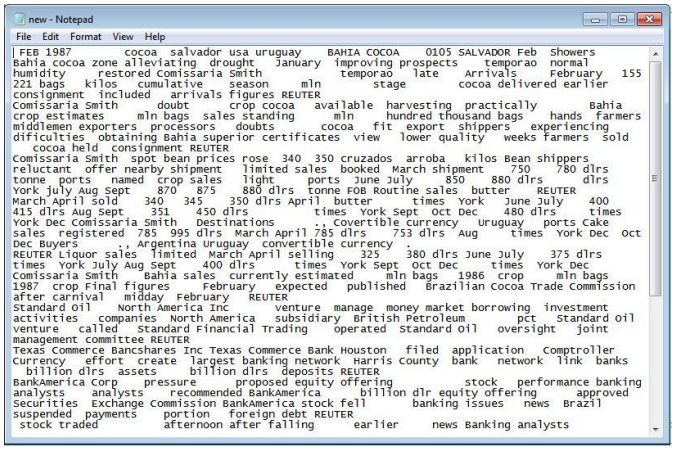

Figure 4: After Stop words removal

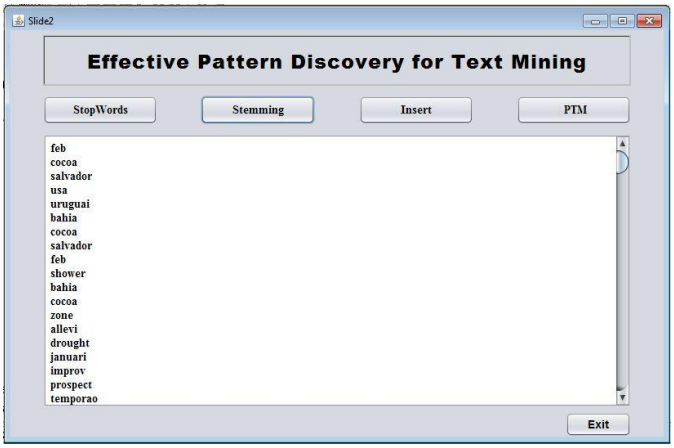

Figure 5: After Stemming

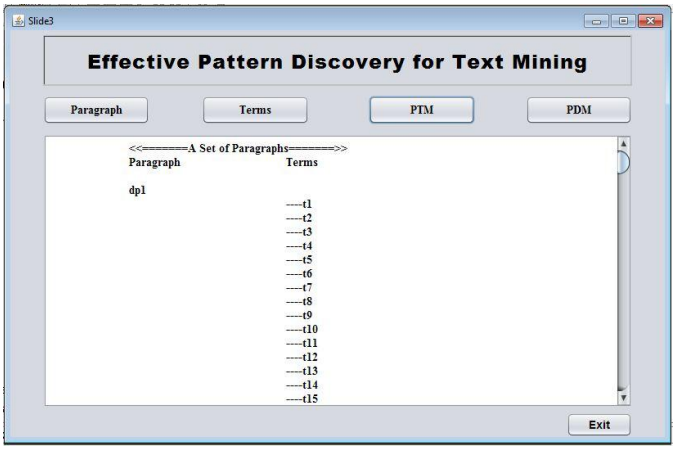

Figure 6: A Set of Paragraphs

5.3 Pattern Taxonomy Model

In this module, all the documents are split into paragraphs. Each paragraph is considered to beach document. The terms, which can be extracted from set of positive documents

Assume that, document d yields a set of paragraphs PS(d). Let D be a training set of documents. Let T = $\{t_1,t_2,.....,t_m\}$ be a set of terms(or keywords).

Parapgraph	*Terms*
dp_1	$t_1\ t_2$
dp_2	$t_3\ t_4\ t_6$
dp_3	$t_3\ t_4\ t_5\ t_6$
dp_4	$t_3\ t_4\ t_5\ t_6$
dp_5	$t_1\ t_2\ t_6\ t_7$
dp_6	$t_1\ t_2\ t_6\ t_7$

Table 1: A Set of Paragraphs

Frequent Pattern	*Covering Set*
$\{\mathbf{t_3}, \mathbf{t_4}, \mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{\mathbf{t_1}, \mathbf{t_2}\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{\mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_6\}$

Table 2: Frequent Patterns and Covering Sets

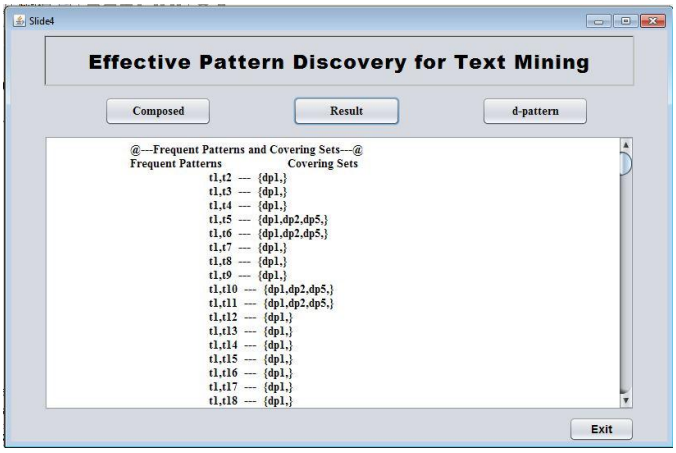

Figure 6: Frequent Patterns and Covering Sets

5.4 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.

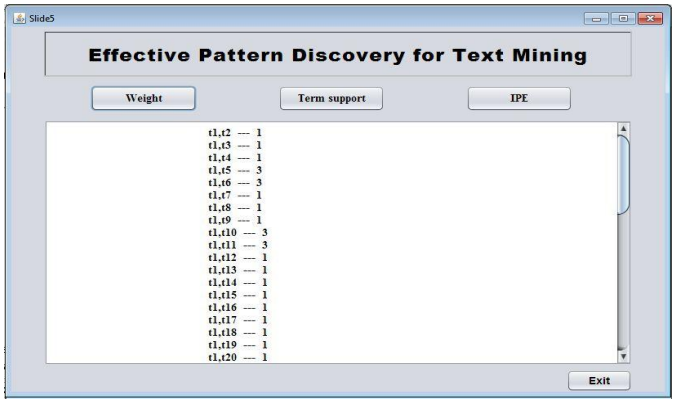

Figure 7: Pattern Deploying

5.5 Pattern Evolving

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.

6. TECHNICAL SPECIFICATIONS

Software Specifications:

Operating System	: Windows 7
Coding Language	: Java
IDE Tool	: NetBeans IDE 8.1
Database used	: MySQL 5.5.
Dataset used	: Reuters21578

Hardware Specifications:

Processor	- Intel i3
RAM	- 4GB
Hard Disk	- 500GB

7. CONCLUSIONS

This paper presents the research on the concept of developing an effective knowledge discovery model (PTM) based on pattern taxonomies. PTM is implemented by three main steps: (1) discovering useful patterns by integrating sequential closed pattern mining algorithm (2) using discovered patterns by pattern deploying; (3) adjusting user profiles by applying pattern evolution. Various mechanisms in each step are proposed and evaluated for fulfilling the PTM model. The latest version of the Reuters dataset is selected and tested by the proposed PTM-based information filtering system. The results show that the PTM outperforms not only several pure data mining-based methods, but also traditional probabilistic and Rocchio methods.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE transactions, vol.24 No. 1, Jan 2012.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

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

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.

[4] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[5] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[8] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[9] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec1.