Effective Pattern Discovery for Text Mining

Dipali C. Sonawane[1], Tejal P. Shirole[2], Kajal D. Patil[3], Priyanka V. Patil[4], , Amol K. Patil[5]

[1] Student, Dept. of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India.
[2]Student, Dept. of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India.
[3]Student, Dept. of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India.
[4]Student, Dept. of Computer Engineering , SSBT COET, Jalgaon, Maharashtra, India.
[5]Student, Dept. of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India.

---***---

Abstract - *In text documents, Data mining techniques have been proposed for mining useful patterns. Mining discovered pattern is still an open issue that how it is effectively use and update, In the domain of text. Term-based approaches are adopted by most existing text mining methods. By the problems of polysemy and synonymy they all suffer. An innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, Text Mining presents. using and updating discovered patterns to improve the effectiveness of for finding relevant and interesting information.*

***Key Words*: Text mining, text classification, pattern mining.**

1.INTRODUCTION

Many data mining techniques have been proposed for extracting useful patterns in text documents. Text mining is the discovery of interesting knowledge in text documents. It is a demanding problem to find accurate knowledge in text documents to help users to find what they want. Data mining techniques have been used for text analysis by extracting occurring terms as descriptive phrases from document groups[1].

Many applications, such as market analysis and business management, can benefit from the use of the information and knowledge taken from a large amount of data. Knowledge discovery can be effectively used and update discovered patterns and apply it to the field of text mining.

Data mining is, therefore, an important step in the process of knowledge finding in databases, which means data mining is having all methods of knowledge discovery process and performing modeling phase that is an application of methods and algorithm for calculation of search pattern or models.

1.1 Background

The interesting knowledge is being found by Text mining in text documents. Find correct knowledge in text documents to help users to find what they want, is a challenging issue. Many term-based methods were supplied by Information Retrieval (IR) to solve this challenge, In the beginning. Term based methods include The advantages such as efficient computational performance as well as mature theories for term weighting, which have appeared over the last couple of decades from the IR and machine learning section. Suffer term based methods from the problems of polysemy and synonymy, a word has many meanings means polysemy and many words having the same meaning is called synonymy. Many discovered terms of the semantic meaning of are uncertain for answering what users want.

In the last decade many data mining methods have been proposed. In the field of text, mining is difficult and ineffective using this discovered knowledge. Because of some useful long patterns with high specificity lack in support. From data mining techniques lead to the ineffective performance was derived by misinterpretations of patterns[2]. To overcome the low-frequent and misinterpretation problems for text mining has proposed by an effective pattern discovery technique. Two processes are used by the proposed method. To refine the discovered patterns in text documents, uses pattern deploying and pattern evolving.

1.2 Motivation

In a collection is now popular such search only marginally supports discovery because the user has to determine on the words to look for, While the ability to search for keywords or phrases. Text mining results can suggest "interesting" patterns look at, On the other hand, and the user can then accept or reject these patterns as interesting. Descriptive frequent patterns are taken by pattern taxonomy model by pruning the unimportant ones is present In this research[3]. Based on their repetitions patterns are sorted.

1.3 Related Work

In the last decade many data mining methods have been proposed. In the field of text, mining is difficult and ineffective using this discovered knowledge. Because of some useful long patterns with high specificity lack in support[4]. From data mining techniques lead to the ineffective performance was derived by misinterpretations of patterns. To overcome the low-frequent and misinterpretation problems for text mining has proposed by An effective[5] pattern discovery technique. Two processes are used by The proposed method. To refine the discovered patterns in text documents, uses, pattern deploying and pattern evolving.

2. LITERATURE SURVEY

Mining Closed Sequential Patterns in Large Sequence Database V. Purushothama Raju1 and G.P. Saradhi Varma ,2015. Information Retrieval using Pattern Deploying and Pattern Evolving Method for Text Mining , ishakha D. Bhope, Sachin N. Deshmukh,2015.

Text mining is nothing but data mining, as the application of algorithm as well as procedure from the machine learning and statistics to text with aim of finding important pattern, Whereas data mining associates in the corporate world because that's where most databases are, text mining give assurance to move machine learning technology out of the companies and into the home" as an more necessary Internet adjunct as web data mining (Hearst, 1997). Laender, Ribeiro-Neto, da Silva, and Teixeira (2001) give a current review of web data extraction tools

3. PROPOSED SYSTEM

In the given proposed system is data from the .text file is taken and analysis is done on it. For analyzing those data the unigram extraction technique is used[6]. For this first preprocessed the dataset, In preprocessing two method is done such as stopword removal and stemming after that extracted the dataset that have use PTM algorithm and then Pattern deployment method applay in which repeatative term of pattern get result.

The process of knowledge discovery may consist of following:

A. Information Selection

B. Information Preprocessing

C. Pattern Taxonomy Model

D. Pattern Deploying

E. Pattern Evaluation.

A. Information Selection

In this module, to load the list of all documents. The user to repossess one among the documents. This document is given to next method. That method is preprocessing.

B. Text Preprocessing

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

1)stop word removal 2)stemming

Stop words are words which are remove natural language data. Stemming is the procedure for reducing words to their stem base form. It normally a written word forms.

C. Pattern Taxonomy Process

In this module, the documents are divided into paragraphs. Each paragraph is considered to be each document. In each document, the set of terms are removed. The terms, which can be removed from a set of good documents.

D. Pattern Deploying

The found patterns are described. The d-pattern algorithm is used to find all patterns in positive documents are combined. The term support is calculated by all terms in d-pattern. Term support means the weight of the term is evaluated[7].

3.1 Pattern Taxonomy Model:

a. All documents are divided into paragraphs. So a given document d yields a set of paragraphs PS(d). Let D be a training set of documents, which consists of a set of good documents, D+; and a set of bad documents, D-. Let T = { t1;t2;...; tm} be a set of terms (or keywords) which can be extracted from the set of positive documents, D+.
b. Representations of Closed Patterns It is complicated to derive a method from applying discovered patterns in text documents for information filtering systems. Let p1 and p2 be sets of term number pairs[8]. p1 p2 is called the composition of p1 and p2.
c. Let DP be a set of d-patterns in DP, and p belongs to DP be a d-pattern. We call p(t) the complete support of term t, which is the number of patterns that carry in the corresponding patterns taxonomies. In order to effectively deploy patterns in dissimilar taxonomies from the dissimilar positive documents.

d. To increase the efficiency of the pattern taxonomy mining, an algorithm, Sequential Pattern Mining, was proposed into search all closed sequential patterns, which used the well-known Apriori property in order to minimize the searching space.

e. Algorithm(PTM) shown in Algorithm describes the training process of finding the set of d-patterns. For each positive document, the SPMining algorithm is first called in step 4 giving rise to a set of closed sequential patterns SP. In Algorithm, complete discovered patterns in a positive document are composed into a D pattern giving rise to a set of d-patterns DP in steps 6 to 9, term supports are calculated based on the normal forms for all terms in d patterns. Let m =|T| be the number of terms in T.

3.2 Implementation

The figure 2 depicts the process flow of a system that consists of loading document for pre- processing by users preference. The text preprocessing, in which the repossess document is passed through two processes such as stop word removal and text stemming. In first process words which are filtered out prior to, or after, processing of natural language data is called as stop words.

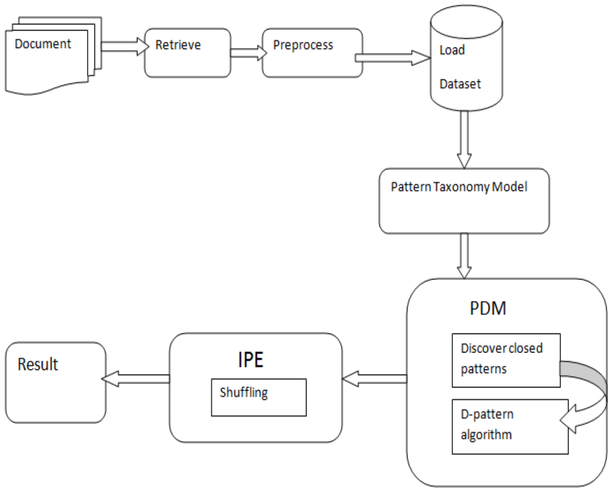

Fig -1 : System Architecture of Proposed System

The second process for reducing inflected (or sometimes derived) words to their stem base or root form called as stemming. In pattern taxonomy process, the documents are divided into paragraphs and are considered as a separate document from which set of terms are removed are called the patterns[9]. In pattern deploying the discovered - patterns are summarized using the d-pattern algorithm. The pattern evolving process is used to determine the noise patterns in documents. In which sometimes, the system falsely identified the bad document as a good. So, noise is occurred in good documents, these noised patterns named as a lawbreaker and if partial conflict lawbreaker contains in good documents, the reshuffle process is applied.

4. RESULT

The proposed system is get implemented in Java language. It takes .text document as input and after applying preprocessing, Pattern Taxonomy Model and Pattern Deployment Model It gives frequent itemset as result i.e pattern that appear in data set frequently.

Table -1: Document Testing

Sr. No.	Characteristic	Dataset 1 value	Dataset 1 value
1.	Load number of document.	1	1
2.	Number of paragraph is in document.	7	7
3.	Number of term required to assign.	3	2
4.	Number of frequent item.	2	2
Result		66%	100%

5. CONCLUSIONS

Text Mining presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

REFERENCES

[1] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,"Effective Pattern Discovery for Text Mining", IEEE Transaction on Knowledge and data engineering, Vol. 24, no. 1, January 2012.

[2] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[3] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[4] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

[5] R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[6] Bharate Laxman1, D.Sujatha, "Improved Method for Pattern Discovery in Text Mining," International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308, Oct-2013.

[7] 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.

[8] V. Purushothama Raju1 and G.P. Saradhi Varma," Mining Closed Sequential Patterns in Large Sequence Databases," International Journal of Database Management Systems (IJDMS) Vol.7, No.1, February 2015.

[9] T. A. Pawar and N. D. Karande ,"Effective Pattern Discovery for Text Mining Using Pattern Based Approach ,"International Journal of advance Research in Computer Science and Management Studies,Volume 2,Issue 9,September 2014.

BIOGRAPHIES

Student, Shram Sadhana Bombay Trust College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India.

Student, Shram Sadhana Bombay Trust College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India.

Student, Shram Sadhana Bombay Trust College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India

Student, Shram Sadhana Bombay Trust College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India.

Student, Shram Sadhana Bombay Trust College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India.