ISSN: 2321-7782 (Online)

Volume 2, Issue 9, September 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Effective Pattern Discovery for Text Mining Using Pattern Based Approach

T. A. Pawar[1]
Assistant Professor
Dept. of CSE
Bharati Vidyapeeth's College of Engg
Kolhapur – India

N. D. Karande[2]
Assistant Professor
Dept. of CSE
Bharati Vidyapeeth's College of Engg
Kolhapur – India

Abstract: In text documents, a significant number of data mining techniques have been proposed for mining useful patterns. But there are some questions; how to effectively use and update discovered patterns is still an open research issue in the area of text mining. Most of existing text mining methods uses term-based approaches but, still they all suffer from the problems of polysemy and synonymy. This paper focuses proposed system implements an effective pattern discovery technique which includes the process of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information for text mining.

Keywords: text mining; stopword; text stemming; patterns.

I. Introduction

A significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms. 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 [1]. Most text mining methods use the keyword based approaches, whereas other choose the phrase based technique to construct a text representation for a set of documents. It is believed that the phrase-based approaches perform better than the keyword-based ones as it is considered that more information is carried by a phrase than by a single term. Although phrases carry less ambiguous and more concise meaning than individual words, the likely reasons for the discouraging performance include: phrases have inferior statistical properties to terms, they have low frequency of occurrence and there are large numbers of redundant and noisy phrases among them [2]. In order to solve above mentioned problems, new system architecture focuses 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. Literature Survey and Proposed System

Text mining is the technique that helps users to find a useful information form a large amount of text data. It is therefore crucial mining model should be retrieve the information that user require with relevant efficiency. In existing, Information Retrieval (IR) provided many term-based methods to solve this challenge. The term-based methods suffer from the problems of polysemy and synonymy. The polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. In proposed to use pattern or phrase-based approaches should perform better than the term-based approaches. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents [3].

The existing system uses term-based approach to extracting the text. In Term-based ontology methods are providing some text representations (for example, hierarchical clustering [6] [7] is used to determine synonymy and hyponymy relations between keywords) and pattern evolution technique is used to improve the performance of term-based approach. The limitation of the term-based approach is it suffered from the problems of polysemy and synonymy and also a term with higher value could be meaningless in some d-patterns (some important parts in documents).

The proposed system is an effective pattern discovery technique, is discovered. It evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns that solves misinterpretation problem. The system 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. In training phase, the proposed model first calls algorithm pattern taxonomy model (PTM) [4] [5] to find d-patterns in positive documents and in testing phase; it evaluates weights for all incoming documents. The incoming documents then can be sorted based on these weights [1]. The proposed approach can improve the accuracy of evaluating term weights because of the discovered patterns are more specific than whole documents. To avoiding the issues of phrase-based approach we considered the pattern-based approach and the pattern mining techniques can be used to find various text patterns.

III. System Architecture

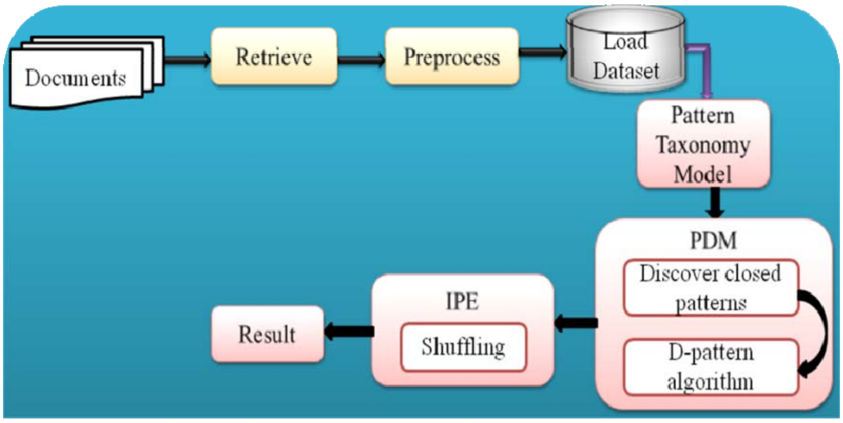

Fig.1 System Architecture

The Fig.1 depicts the process flow of system that consists of loading document for preprocessing by users preference. The text preprocessing, in which the retrieved document is passed through two processes such as stopword removal and text stemming. In first process words which are filtered out prior to, or after, processing of natural language data are called as stopwords. 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 split into paragraphs and are considered as a separate document from which set of terms are extracted are called the patterns. In pattern deploying the discovered patterns are summarized using d-pattern algorithm. The pattern evolving process is used to identify the noisy patterns in documents. In which sometimes, the system falsely identified negative document as a positive. So, noise is occurred in positive documents, these noised patterns named as offender and if partial conflict offender contains in positive documents, the reshuffle process is applied [1].

IV. Experimental Work

The experimental work carried out in which considers a dataset having list of documents. First selection of document then passes it for text preprocessing where stopword are calculated and displayed. The stemming finds root words from document which is results of stopword removal and classifies the document in paragraph format. Apply the pattern taxonomy process, in which terms are extracted and synonymies as well frequency calculation is done for every keyword. The discovered patterns from removal of sub pattern done by closed sequential pattern and the D-pattern calculated from closed sequential patterns. To search the document the keyword is inserted, after processing the search result containing list of documents are displayed.

V. Conclusion and Future Work

The system works satisfactorily for effective pattern using pattern based approach for text mining resulting in d-pattern which is nothing but list of terms calculated by performing set of operation on list of closed sequential patterns. The system performs stopword removal, streaming and frequency count successfully.

The experiment that conducted successfully, which is the product of different modules implemented accordingly text preprocessing, pattern taxonomy process, pattern deploying. The obtained results in d-pattern are nothing but the list of terms calculated by performing the set of operations on list of closed sequential patterns. The system is limited to work on, inner pattern evolution as unable to identify the negative document in dataset. The operation was only performed on single folder of current dataset which considers only xml file documents for obtaining the patterns.

The future scope is to work on inner pattern evolution in which it consider the negative document term set for identifying falsie positive document and elaborate the system to work properly for handling real file system.

References

1. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, IEEE transactions, vol.24 No. 1, Jan 2012.
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 Modelon the Web and Its Application in Jobagent,” Knowledge-BasedSystems, vol. 13, no. 5, pp. 285-296, 2000.

Author(s) Profile

Mrs. T. A. Pawar, received the B. E. degree in Computer Science and Engineering from Dr. D. Y. Patil College of Engineering, Kolhapur, India in 2007. She is doing her M.Tech in Computer Science at Jawaharlal Nehru University, Hyderabad, India. From 2010 to till date she is working as Assistant Professor at Bharati Vidyapeeth College of Engineering, Kolhapur, India. She has published various papers in area of Network Security and Data Mining.

Mr. N. D. Karande, received the B.E. degree in Computer Science and Engineering from Bharati Vidyapeeth College of Engineering, Kolhapur, India in 2006. He received the M.Tech degree in Computer Science and Technology from Shivaji University, Kolhapur, India in 2010. From 2008 to till date, he is working as Assistant Professor at Bharati Vidyapeeth College of Engineering, Kolhapur, India. He has published various papers in the area of Database Indexing, Security and Natural Language Processing.