IJETCR Open Access

ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2; Issue 6; Page No. 120-123

A Study on Effective Pattern Discovery Using Text Mining

Shivani Gupta [1], B.P. Vasgi [2]

[1]P.G. Student, Department of IT, Sinhgad College of Engineering, Pune, Maharashtra, India[1]

[2]Assistant Professor, Department of IT, Sinhgad College of Engineering, Pune, Maharashtra, India[2]

shivanigupta2608@gmail.com

ABSTRACT

A significant number of data mining techniques are proposed for mining the useful patterns in text documents. But in the area of text mining, the question; how to update and use these discovered patterns is open research issue. Many of the text mining methods use the term based approaches. The term based approaches faces the problem of polysemy and synonymy. This paper focuses on the implementation of an effective pattern discovery technique. The process of pattern deploying and pattern evolving is included in the pattern discovery techniques. These techniques improve the effectiveness of the discovered patterns. The discovered patterns are used for finding relevant and interesting information for text mining.

Key Words: Patterns, pattern deploying, pattern evolution, text mining, stemming.

I. INTRODUCTION

Text mining is the process of discovering an interesting knowledge from the text documents. It is very difficult task to find the knowledge in the text documents which will be helpful for the users to find what they want. To overcome this difficulty, information retrieval provided many term based methods such as rough set models [23], Rocchio and probabilistic models [4], BM25 and support vector machine (SVM) [34] based filtering models.BM25 and support vector machine (SVM) [34] based filtering models. The advantages of term based methods are that it includes computational performance and some theories for term weighting. These theories have emerged from the IR and machine learning communities.

In order to perform different knowledge tasks number of data mining techniques have been proposed. Association rule mining, maximum pattern mining, sequential pattern mining, closed pattern mining, frequent itemset mining are included in these techniques. These techniques are generally proposed for developing efficient mining algorithms. There are number of data mining approaches used for the text mining such as keyword based approach, phrase based approach etc. These approaches are used to prepare a text representation from a large set of documents. Generally the phrase based approach performs better than the keyword based approach. Phrases carry much more and specific information rather than a single keyword or single term. But phrase based approach also have some drawbacks: 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]. To overcome the above stated problem a new system id proposed which focuses mainly on the knowledge discovery and the effective use of the discovered patterns and applying it to the field of text mining.

II. 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 means a single word which has multiple meanings and synonymy means multiple words which have the same meaning. 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 [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 technique are very useful in finding the text patterns.

III. SYSTEM ARCHITECTURE

The Fig.1 depicts System Architecture of the pattern based approach.

Preprocessing:

In preprocessing module, the document has to go through two important processes.

a) Stopword Removal:

In this process the whole document is scanned and all the stopwords are filtered out. The words like "is", "or" and "she", "he" etc are known as stopwords.

b) Stemming:

In this process the derived words are cut down to their stem or base root words. For example learning is converted into learn, filtering is cut down to filter. Removal of "ing","ed","es" etc comes under stemming process.

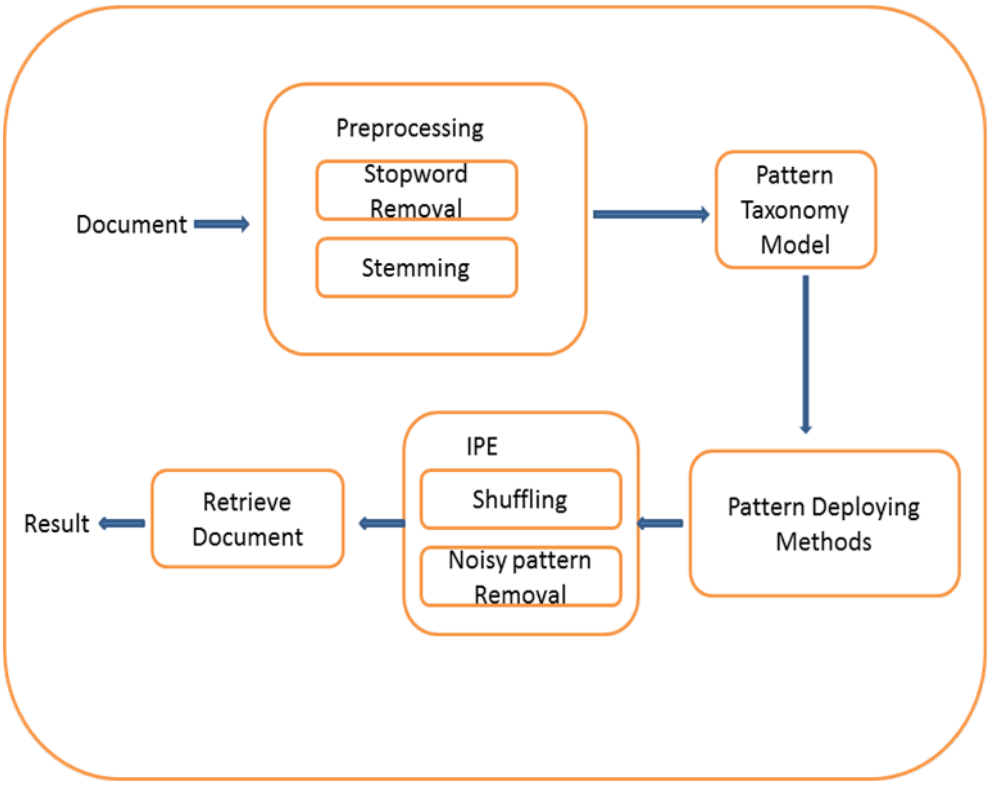

Figure 1: System Architecture

Pattern Taxonomy Model:

In pattern taxonomy model, the uploaded document is first scanned and the whole document is converted into a set of paragraphs. These paragraphs are treated as a separate document. From these documents number of set of terms are extracted. These extracted terms form a particular pattern. The discovered patterns are checked with the help of d- pattern algorithm.

Inner Pattern Evolution:

In inner pattern evolution module, the noisy patterns available in the documents are identified. Noisy patterns refer to some false documents or negative documents which are considered as positive by mistake. That's why some kind of noise is there in the positive document collection. These patterns are generally referred as offenders. The reshuffle process is applied on these offenders to remove the noisy and redundant patterns.

IV**. Hardware and Software Requirements:**

Hardware Requirements:

- RAM: 2GB
- Hard Disk: 500 GB
- Processor : i3
- OS: XP, Win 7.

Software Requirements:

- Netbean 7.1
- JDK 1.7

V**. Description of the existing system**

Sr. No	Title	Author name	Work description	Drawbacks
1	Automatic Text Categorization and Its Application to Text Retrieval.[9]	Wai Lam et.al	Demonstrated the effectiveness of automatic text categorization approach and investigated its application to text retrieval.	The instance based categorization method used for automatic TC is limited and less feasible.
2	Statistical Phrases in Automated Text Categorization [7]	Stan Matwin et.al	Investigated the usefulness of n-grams in text categorization independently of any specific learning algorithm.	The Results were dependent on: a) Chosen heuristics. b) Chosen classifier learning.
3	Improving the retrieval of information from external sources.[8]	Susant Dumais	Described a statistical method called latent semantic indexing, and improved the retrieval performance by up to 30%.	Users faced difficulty in stating their requests in a way that leads to appropriate query placement.
4	Mining association language patterns using a distributional semantic model for negative life event classification[11]	Liang-Chih Yu et.al	A combination of both supervised data mining algorithm and unsupervised distributional semantic model is presented in the proposed framework. This framework is used to discover association language patterns.	There are vast amount of the data that is built in the informal terms. Hence the text categorization is less suitable for the informal English language that is based on web content.
5	Mining Association rules from unstructured documents.[10]	Hany Mahgoub et.al	Author presented a system for discovering association rules from collections of unstructured documents called EART (Extract Association Rules from Text)	The work depends on the analysis of the keywords in the extracted association rules through the co-occurrence of the keywords in one sentence in the original text and the existing of the Keywords in one sentence without co-occurrence. Weight of the term evaluated can be misleading sometimes.

VI. Conclusion and Future Work

The proposed system performs satisfactorily for effectively discovering the patterns by using the pattern based approach. The pattern based approach is used for text mining which results in d patterns. D patterns are the just the list of number of terms which are calculated by performing some set of algorithms or operations on list of sequential patterns. The system Performa all the operations like frequency count, stemming stop word removal etc. successfully.

Number of experiments were conducted which were successful. Experiments basically included the implementation of all the modules like preprocessing, inner pattern evolution and pattern taxonomy process etc. D patterns are the just the list of number of terms which are calculated by performing some set of algorithms or operations on list of sequential patterns. Inner pattern evolution process was necessary to avoid the negative document set in the list of patterns. Pattern based approach is better than the phrase based approach as it overcomes all the drawbacks of the phrase based approach.

The future scope is to work on the retrieval of the relevant documents by based on the discovered patterns. And elaborate the system more and more to work successfully in handling the real file system

VI. 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. S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
9. W. Lam, M.E. Ruiz, and P. Srinivasan, “Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowl-edge and Data Eng.,vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
10. Hany Mahgoub, "Mining Association rules from unstructured documents", World Academy of Science, Engineering and Technology, Vol.20, No.1, pp.1-6, 2006.
11. Liang-Chih Yu, Chien-Lung Chan, Chao-Cheng Lin,I-Chun Lin,”Mining association language patterns using a distributional semantic model for negative life event classification”,Journal of Biomedsical Informatics,Vol.44,no. 4, August, 2011.
12. 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), 1998.
13. T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization,” Proc. 14th Int’l Conf. Machine Learning (ICML ’97),pp. 143-151, 1997