International Journal of u- and e- Service, Science and Technology
Vol.9, No. 7 (2016), pp.197-202
http://dx.doi.org/10.14257/ijunesst.2016.9.7.20

Data Mining Techniques Based on Effective Pattern Discovery

S. Suresh Babu[1], Vahiduddin Shariff [2], CH. M. H. Saibaba[3]
and Debnath Bhattacharyya[4*]

[1]Department of Information Technology, VFSTR University,
Vadlamudi-522213, Guntur, India
[2]Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru, West Godavari district
[3]Department of Computer Science and Engineering, KL University,
Vadddeswaram-522502, Guntur, India
[4]Department of Computer Science and Engineering,
Vignan's Institute of Information Technology,
Visakhapatnam-530049, India
[1]suresh.satukumati@gmail.com, [2]shariff.v@gmail.com,
[3]saibaba.ch77@kluniversity.in, [4] debnathb@gmail.com*

Abstract

The extraction of similar features based on quality is called pattern discovery to the huge number of terms, phrases and noise. Identifying the better pattern discovery is the major problem to extract the accurate information from the text documents because of the noise and unwanted data present in the text documents. In this paper, pattern discovery is used to find the frequent item sets and reducing the noise from text documents and implement the advanced pattern discovery approach. In this paper, for implementation we use .txt files with unstructured data to find the efficient patterns.

Keywords: *Mining, textual content, term based techniques*

1. Introduction

Because of the fast growth of digital information made available in current years, knowledge discovery and data mining have attracted a superb deal of attention with an approaching want for turning such data into useful facts and know-how. Many programs, which include marketplace evaluation and business management, can benefit through using the information and understanding extracted from a big amount of records.

Knowledge discovery can be regarded as the technique of nontrivial extraction of records from large databases, facts that is implicitly offered within the records, previously unknown and doubtlessly useful for customers. Records mining are therefore a crucial step in the manner of expertise discovery in databases. Inside the beyond decade, a widespread number of records mining strategies were presented in order to carry out unique expertise obligations. Those strategies encompass association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed sample mining. Most of them are proposed for the motive of developing green mining algorithms to find specific patterns inside an inexpensive and desirable time body. With a huge range of styles generated through using information mining techniques, the way to efficiently use and update those styles remains an open studies trouble. In this paper, recognition at the development of an information discovery version to correctly use and replace the discovered styles are proposed and apply it to the sector of text mining.

[*] Corresponding Author

ISSN: 2005-4246 IJUNESST

Textual content mining is the discovery of interesting know-how in textual content documents. It's miles a difficult issue to locate accurate understanding (or functions) in textual content documents to help customers to locate what they need. inside the starting, records Retrieval (IR) provided many term-based totally techniques to remedy this project, which include Rocchio and probabilistic models, tough set models, BM25 and support vector device (SVM) based totally filtering models. The blessings of term based methods include green computational overall performance in addition to mature theories for time period weighting, which have emerged over the past couple of decades from the IR and machine gaining knowledge of communities. however, term based methods be afflicted by the issues of polysemy and synonymy, wherein polysemy approach a word has a couple of meanings, and synonymy is more than one words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what customers want.

There are essential troubles regarding the effectiveness of sample-based totally process: low frequency and misinterpretation. Given a particular subject matter, a especially common pattern (generally a short pattern with large assist) is generally a general pattern, or a specific pattern of low frequency. If we lower the minimal support, loads of noisy styles might be located. Misinterpretation method the measures used in pattern mining (*e.g.*, "help" and "self belief") come to be not suitable in using found styles to answer what customers need. The tough problem consequently is the way to use found patterns to appropriately examine the weights of beneficial capabilities (know-how) in text files.

So as to remedy the above paradox, this paper offers an effective sample discovery method, which first calculates located specificities of patterns after which evaluates term weights consistent with the distribution of phrases in the observed patterns rather than the distribution in files for fixing the misinterpretation problem. It also considers the have an impact on of styles from the poor education examples to locate ambiguous (noisy) patterns and try and lessen their impact for the low-frequency problem.

The manner of updating ambiguous patterns can be referred as sample evolution. The proposed approach can enhance the accuracy of comparing term weights because found styles are more precise than complete documents.

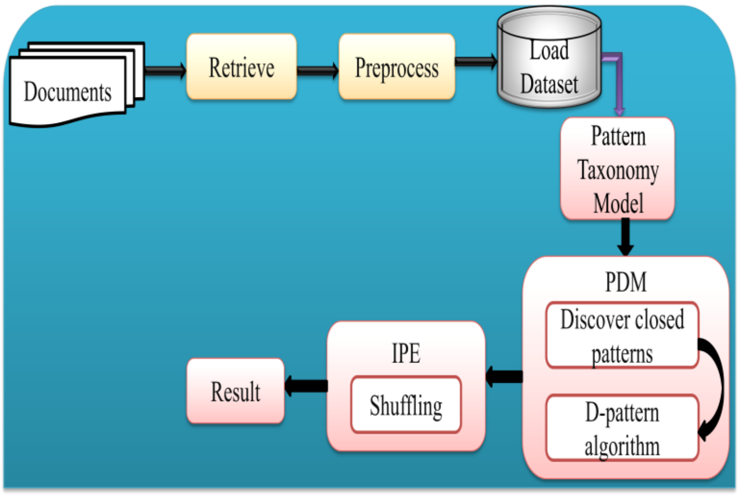

Figure 1. System Architecture

2. Existing System

In the existing work to extract the text we use term-based approach is implemented. These methods represent the term based ontology. To determine the synonymy and hyponymy relations between keywords Hierarchical is used. To improve the performance of the term based we use pattern approach.

2.1. Disadvantages of Existing System

The term-based approach is suffered from the problems of polysemy and synonymy. A term with higher value could be meaningless in some d-patterns (some important parts in documents).

3. Proposed System

An advanced pattern discovery technique is discovered. Appraise specificities of patterns and then appraises term weights according to the distribution of terms in the discovered patterns. Solves falsify Problem. Training the samples to find the noisy patterns and influence to reduce the low-frequency problem. In this pattern evolution, the process of updating ambiguous patterns is referred. We can identify the improvement by using proposed approach by evaluating term weights because discovered patterns are more specific than whole documents.

There are two modules in System they are 1) Training and 2) Testing.

1) In training module, the d-patterns in the positive documents (pd) divide on min sup are identified, and evaluates term supports by deploying d-patterns to terms.

2) In testing module, it will test the noise negative documents in D based on experimental coefficient. Based on the weights the incoming documents are sorted.

3.1. Advantages of Proposed System

To improve the performance of the evaluating term weights by using proposed system. From all the documents the identified documents are more important. To avoid the issues of phrase-based approach using pattern-based approach. To find out various text patterns we use pattern mining techniques.

4. Proposed Work

Proposed work was accomplished in the following Modules:

4.1. Loading Document

- Load all the list of documents.
- User will search for one document.
- Searched document will move to next process.

4.2. Text Preprocessing

The searched document process is done in this module. There are two types of process is done.

1) Stop words removal.

2) Text Stemming.

Stop words will be words which are sifted through before, or in the wake of, handling of characteristic language information.

Stemming is the procedure for decreasing arched (or at times inferred) words to their stem base or root structure. It for the most part a composed word forms.

4.3. Pattern Taxonomy Process

In this module, the records are part into paragraphs. Each paragraph is thought to be every report. In every record, the arrangement of terms are separated. The terms, which can be extricated from set of positive reports.

4.4. Pattern Deploying

The positive documents are extracted from set of terms. The d-design calculation is utilized to find all examples in positive archives are made. The term backings are computed by all terms in d-design. Term support implies weight of the term is assessed.

4.5. Pattern Evolving

In this module used to distinguish the noisy patterns in records. Sometimes, framework dishonestly recognized negative record as a positive. So, noise is happened in positive record. The noised design named as wrongdoer.

If incomplete clash guilty party contains in positive documents, the reshuffle process is applied.

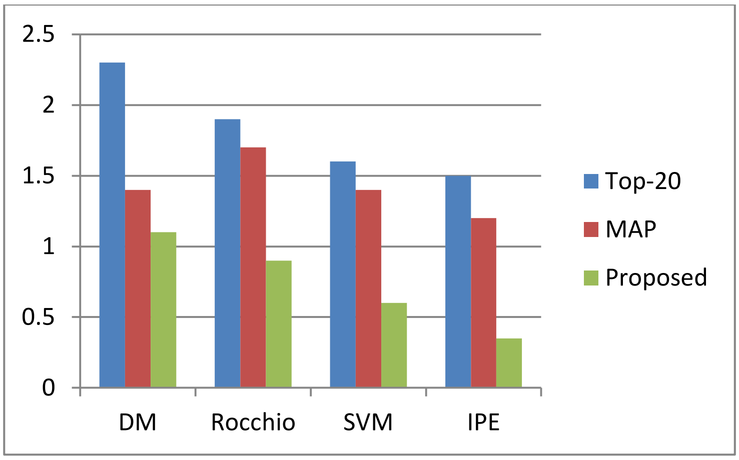

Figure 2. Comparison of Proposed and Other Major Models in Three Measures for the 100 Topics

5. Conclusion

In data mining, text mining is the most important technique to find the frequent patterns from the various .txt files or csv files and from various huge data sources. Though there are number of mining techniques like association rule mining, common item set mining, sequential sample mining, most sample mining, and closed sample mining. Still there is a lack (*i.e.*, low frequency) of identifying the similar patterns by using above data mining techniques. In this proposed work, we have mainly focus on finding and search the efficient pattern mining information from large datasets. In proposed technique we can take input file .txt then we apply various algorithms such as PTM, PDM, D-Pattern, IPE for Shuffling Inner pattern & display expected output. The proposed system implements two processes, pattern deploying and pattern evolving, to extract the efficient discovered patterns in text documents. The experimental results shows the performance of the proposed system based on outperforms no longer most effective different natural statistics mining-primarily based strategies and the concept based model, but also time period-based modern fashions, consisting of BM25 and SVM-based totally fashions.

References

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, **(1999)**.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94) , **(1994)**, pp. 478-499.

[3] 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)**, pp. 2-11.

[4] R. B. Yates and B. Ribeiro-Neto, "Modern Information Retrieval", Addison Wesley, **(1999)**.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov/ pubs/trec11/papers/kermit.ps.gz, **(2002)**.

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

[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] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, **(1995)**, pp. 273-297.

[9] S. T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, **(1991)**, pp. 229-236.

[10] J. Han and K. C. C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, (**2002)**, November, pp. 64-70,