A New Approach for Extraction of Pattern Frames in Text Mining

B. Sankara Babu
Associate Professor
Department of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad

K. Rajasekhar Rao, PhD
Professor
Department of CSE
Koneru Laxmaiah University
Guntur

P.Satheesh, PhD
Associate Professor
Department of CSE
Maharaj Vijayaram Gajapathiraj college of Engineering
Vizianagaram

ABSTRACT

Due to the rapid growth in World Wide Web and data availability, text mining has become one of the most important fields in data mining. Text mining refers to the technique which is useful to find the information from a huge volume of digital documents. Many existing text mining methods follow the term based approaches. Pattern evolution methods are employed to perform the same concept of tasks .This paper presents a new approach for extraction of the pattern frames in text mining.

Keywords - text mining, pattern frames, information

1. INTRODUCTION

Data mining is a process of extracting knowledge from massive collection of data. It refers to a way of finding significant and useful information from data. Organizations that make use of data mining techniques are benefited in their corresponding business area by identifying the significant trends and anomalies that were not possible to be detected by a human analysis. Data mining techniques are used to extract the information that is potentially useful for users[8]. Text mining is interpreted as finding the interesting knowledge in text documents. It is a challenging issue to find the exact information that the users need. The traditional Information Retrieval (IR) has the same objective of retrieving relevant documents and filtering the non- relevant documents. Key-word based method was adopted for that. The term-based methods have been provided with term weights. The term based methods came across with the problem of polysemy and synonymy. People thought that phrase based methods could perform better than term-based methods[7]. But it has its own disadvantages includes low occurrence and inferior statistical properties to words. The objective of the current work is to propose an approach for finding the useful patterns in the text documents.

2. RELATED WORK

The popular tf*idf (term factor*inverse document factor) weighting scheme is used for text representation in Rocchio classifiers [14]. This method is based on the frequencies of individual terms in the document and frequencies of words in entire collection. There is another representation of a document in the form of binary vector. Words are represented considering their existence in the document [1].In this representation it is possible to use variety of categorical data clustering algorithms on binary representation..

The problem with this bag of words approach is selection of limited number of features from a large set of terms. Usually a term with high weight may be a general term. But it might not be relevant to the user interestingness.

For example in the search string “Applications of computers in the field of Education” the terms education and computer are general terms which are likely to have more weight. But when these terms are associated with the prefix or post fix terms in the search string are going to be the user’s interestingness measures.

Some data mining techniques that have been used with descriptive phrases [10] showed no significant improvement in the efficiency in mining.

One alternative method for those two is pattern mining. A number of algorithms such as Apriori[5],Spade[6] have been proposed. The main objective of that algorithms are to discover useful and interesting patterns from huge volume of data[2]. The discovered patterns can be utilized in many ways. By evaluating the term weights in the discovered patterns the efficiency can be improved.

The proposed model describes how to extract patterns from text documents .Each document comprises a certain number of paragraphs. First the documents are cleaned. We apply standard Porter-Stemmer algorithm for stop words removal and stemming. Stop words are frequent words that carry no information for example words like ‘is’,’this’, ‘of ‘ etc .By word stemming we mean the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as call, called, and calling. The porter -stemmer is a well known algorithm for this task.

3. METHODOLOGIES

3.1 Definitions*:*

*(*a) Document set*:* Let D be a set of documents it comprises a set positive and negative documents. For mining we consider the positive documents only.

$D=D^{+}\ U\ D^{-}$

(b)Paragraph set*:* Each document consists of a finite number of paragraphs. It is called as a paragraph set.

d={tp1,tp2,…,tpm}

(c) Term set*:* T={t1,t2,…,tn} is a set of terms extracted from positive documents.

(d) Pattern*:* An ordered set of terms in a document is called as pattern.

(e) Covering set*:* A set of paragraphs which includes the given pattern in a document. Covering set of a pattern X is denoted by [X].

The number of paragraphs containing a particular pattern is called the absolute support of the pattern.

$sup_a(x)=[x]$

dividing the absolute support of a pattern with total number of paragraphs gives the the relative support of the pattern

$sup_r(x)=[x]$/total number of paragraphs;

Pattern Taxonomy Model : Having been preprocessed the documents will be given to the PTM. This will split the document into paragraphs. Each paragraph is treated as individual document and data mining methods are applied. Consider the following Table.

Fig.1.Paragraphs in document

Text paragraph	Term sequence
tp1	t1 t2
tp2	t3 t4 t6
tp3	t3 t4 t5 t6
tp4	t3 t4 t5 t6
tp5	t1 t2 t6 t7
tp6	t1 t2 t6 t7

The process of pattern taxonomy model can be described as the sequential steps of preprocessing followed by splitting the documents into paragraphs and finding the frequent patterns and then the closed sequential patterns. For d-pattern mining an efficient algoritm was proposed by Ning Zhong Yufeng LI and shang Thang wu[5]

Extraction of frequent sequential patterns is the process of appending the terms to existing sequence of terms in the same order as they appear in the document.

First a term t2 is found in the document and then t3 is found then the pattern will be t2, t3 and later on t1 is found then it will be appended and the pattern would be t2 t3 t1.

Algorithm: for finding sequential patterns

Input: set of paragraphs, min_sup

Output: set of frequent patterns

1. Take each paragraph from the set of paragraphs
2. Search for each term in the term set
3. Append the subsequent terms that are found to the term that has been previously found
4. Repeat the steps from 1 to 3 for all the paragraphs of a document.

Once the sequential patterns are obtained the effacious patterns has been calculated. Usually the interestingness of a particular user will come to known only by observing the search string given by the user. In general the first few terms of the search string are going to be very crucial for searching. After removal of wh question words, prepositions and articles in the search string given by the user, there is chance of having only root words. Among the root words the first three or four words are very important for carrying the search process.

After obtaining the frequent sequential patterns now pick up the patterns that starts with term t1 (i.e the first term in the search string) and then find the patterns that have the immediate following terms with t2 ,t3 or t4 etc. It means search for the next three terms after the first term if any of the terms t2, t3, t4 exists in those terms then identify it as an efficacious pattern.

Algorithm: find the efficacious pattern from the sequential patterns

```
Input: sequential patterns
Output: efficacious patterns
1 .EP=∅
2. for each pattern pi in SP do
3. if pi[1]=t1 then
4. search next three terms
5. if(term found is in (t2,t3,t4))
6.  NP =t1®x   {x is in [t2,t3,t4])
    EP=EP U NP
7. if pi[1]=t2 then
8. search next three terms
9. if(term found is in (t1,t3,t4)) then
10. NP =t1®x {x is in [t1,t3,t4])
     EP=EP U NP
11. end
```

Initially there are no efficacious patterns this is indicated by EP=∅. Now taking each pattern in the sequential pattern pick the first term and check whether it is either first or second term in the search list. If it is t1 then consider the consequent three terms and check for any of the three terms t2, t3 or t4 exists in the pattern if so then identify it as efficacious pattern, or else if the first term of the pattern is t2 then consider the next three terms and check for the three terms t1, t3, t4, if any of the terms are found then identify it as an efficacious pattern.

The above algorithm finds out the patterns that are close to the interestingness of the user.

4. RESULTS

In this section results using have been produced basing on the approach proposed as well as using other approaches such as sequential pattern which is part of data mining method based on svm, sequential closed patterns data mining method in scpm.

Table 4.1 shows sequential pattern and top match

S.No	Initial Pattern	Sequential Patterns	Top Match
1	t1	t1, t4,t3	0.45
2	t3	t3, t5,t4	0.35
3	t2	t2,t3,t5	0.54
4	t4	t4,t2,t3,t5	0.44
5	t5	t5,t2,t4	0.38

The above table illustrates the initial pattern, sequential patterns and their top match of a particular document contains different paragraphs. By considering the initial pattern(t1) of a document Sequential patterns are generated by using the Sequential pattern algorithm that is t1,t4,t3 and got top match as 0.45. Similarly different patterns are considered.

Table 4.2 shows sequential closed pattern and top match

S.No	Intial Pattern	Sequential closed Patterns	Top Match
1	t1	t1, t3,t5	0.52
2	t3	t3, t5,t1	0.43
3	t2	t2,t3,t4	0.69
4	t4	t4, t3,t5	0.53
5	t5	t5,t4	0.44

The above table illustrates the Initial pattern, Sequential closed patterns and their Top match of particular document contains various paragraphs. Here the Initial Pattern(t1) is considered from the paragraph, the relevant Sequential closed patterns that is t1,t3,t5 by using the Sequential closed pattern algorithm has been derived and their top match is calculated to be 0.52. Similarly different initial patterns are considered and sequential closed patterns and their top match have been defined.

Table 4.3 shows comparison between sequential, closed sequential and Search Term priority.

S.No	Sequential Patterns	Sequential Closed Patterns	Search Term Priority
1	0.45	0.52	0.49
2	0.35	0.43	0.46
3	0.54	0.59	0.57
4	0.44	0.50	0.53
5	0.38	0.44	0.46

The table 4.3 results represent the top matches of Sequential patterns, Sequential closed Patterns and proposed Search Term Priority approach of particular document contains various paragraphs. Here Search Term Priority, Sequential patterns and the Sequential Closed Patterns have been applied. Priority is given by examining the top matches of Sequential patterns and Sequential closed patterns. In the above table Sequential patterns of particular document is 0.45 and sequential closed pattern is 0.52 and the search term priority was obtained to be 0.49. Better results have been derived after applying various document level constraints.

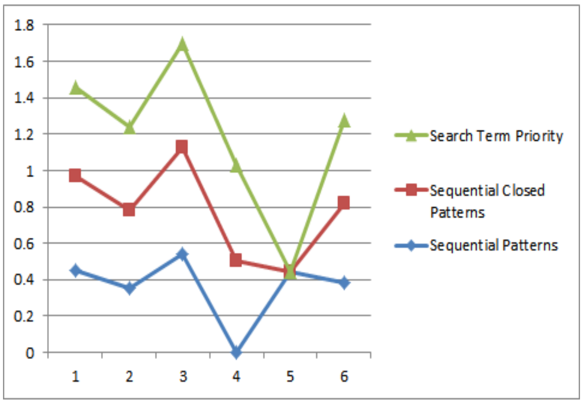

Figure 2. Figure depicting different search patterns

The above graph crystal clears the searching strategies of using different algorithms. It has been clearly shown that the search term priority algorithm has become more prominent in obtaining the top matches out of the documents that have been considered, being derived in terms of paragraphs. This graph clearly illustrates the sharp of searching the key notes which are in general termed as top matching strategies.

5. CONCLUSION

Many techniques in data mining adopted term support methods. These methods worked on term support by ignoring the relationships between the terms. This paper presents a method that finds out the patterns based on the user interestingness measures by taking into consideration of the search terms priority.

6. REFERENCES

[1] N. Zhong, Y. Li, and S.T. Wu. Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, DOI: http://doi.ieeecomputersociety.org/10.1109/TKDE.2010 K.

[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), pp. 478-499, 1994.

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

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

[5] J.S. Park, M.S. Chen, and P.S. Yu, “An Effective Hash-Based Algorithm for Mining Association Rules,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’95), pp. 175-186, 1995.

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

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, “Statisticals 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] M. Seno and G. Karypis, “Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint,” Proc. IEEE Second Int’l Conf. Data Mining (ICDM ’02),pp. 418-425, 2002.

[10] J. Han and K.C.-C. Chang, “Data Mining for Web Intelligence,” Computer, vol. 35, no. 11, pp. 64-70, Nov.2002.

[11] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), pp. 1-12, 2000.

[12] Y. Huang and S. Lin, “Mining Sequential Patterns Using Graph Search Techniques,” Proc. 27th Ann. Int’l Computer Software and Applications Conf., pp. 4-9, 2003.

[13] D.D. Lewis, “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task,” Proc. 15th Ann. Int’l ACM SIGIR Conf. Research and

Development in Information Retrieval (SIGIR ’92), pp. 37-50, 1992.

[14] 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.

[15] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proc. European Conf. Machine Learning (ICML ’98),, pp. 137-142, 1998.

[16] T. Joachims, “Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 200-209, 1999.

[17] W. Lam, M.E. Ruiz, and P. Srinivasan, “Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999. 42 IEEE TRANSACTIONS ON International Journal of Advances in Science Engineering and Technology Volume- 1, Issue- 1.

[18] Words Sequence Pattern Mining Using Pattern Taxonomy Model Knowledge and Data Engineering, Vol. 24, No. 1, January 2012