International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37747-37750
© Research India Publications. http://www.ripublication.com

An Improved Pattern Extraction from Frames in Text Mining

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

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

Dr.P.Satheesh
Associate Professor Department of CSE MVGR College of Engineering Vizianagaram

Abstract
Text-mining refers to the process of extracting interesting and non-trivial information from unstructured text. This is done by many methods such as term based methods and phrase based methods. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy.
In this paper, we propose a model for discovering frequent sequential patterns, phrases, which can be used as profile descriptors of documents. However, it is difficult to use these phrases effectively for answering what users want. Therefore, a pattern taxonomy extraction model is presented which performs the task of extracting descriptive frequent sequential patterns by pruning the meaningless ones.
The experimental results are calculated based on the accuracy, precision and recall rates. The result shows that the pattern-based taxonomy model outperforms the keyword-based methods to represent the documents.

Keywords: Pattern Taxonomy Model, Polysemy, Synonymy, Sequential Patterns.

Introduction

Text mining is the process of discovering the interesting knowledge in text documents. It is a challenging issue to find accurate knowledge in text documents and to help users to find what they want. By using the information extracted from a large amount of data, many applications like market analysis and business management can be benefited. Data mining is therefore an essential step in the process of knowledge discovery in databases, which means data mining is having all methods of knowledge discovery process and presenting modeling phase that is application of methods and algorithm for calculation of search patterns or models. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame.A large number of patterns are generated by using the data mining approaches, how effectively these patterns can be exploited is still an open research issue.
In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering models. The advantages of term-based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term-based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what user wants.

Literature Review

A pattern mining model consists of pattern deploying, inner pattern evalution, fuzzy estimated and similarity based self-generating algorithm for text classification is proposed in [1]. It is one of the extended fuzzy feature clustering algorithms in text classification. The input to the proposed system is a textual document. It is collecting from different sectors such as newspapers, articles, journals, magazines, university details and IT industry. Here the word is extracted from text document. Same words are classified into one group and different words are categorized into another group. After classification, the similarity between different words is calculated. Then the weights for each word are formed automatically in a desired number of categorizations. Four ways of weighting such as paragraph, document, concept and sentence are discussed. The fuzzy algorithm properly deals with the weighting scheme between different features. It is very easy to avoid the low frequency and misinterpretation problem in pattern mining. Some real world experiments are conducted on data sets RCV1 and TREC. It shows excellent improvements and also run faster, reduces storage requirements.
In text mining methods, most of the existing pattern mining techniques go through the problems of lack of accuracy and lack of performance. In [2], the pre-processing step progresses by removing the “noise” word. The next development is the Hidden Markov Models (HMMs), which are used for pattern extraction and classification of input data. HMM calculates the possibility value between noticed events and unnoticed events. This method can improve the accuracy of evaluating term weights and also used to progress the performance for discovering patterns in text for large databases.
An innovative and effective method that extends the random set to accurately weight patterns based on their distribution in

the documents and their terms distribution in patterns is presented in [3] which will find the specific closed sequential patterns (SCSP) based on the new calculated weight.
A Naive Bayesian algorithm used for discovering of patterns is presented in [4], as it is the most appropriate one for classifying positive and negative documents. The usual results will not be in an optimized manner. The prescribed method makes the output arranged in a particular order.
A new approach for pattern discovery technique is proposed in [5] which evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns and solves misinterpretation problems. It considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries to reduce their influence on the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evaluation. The proposed approach improves the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.
An innovative and effective Prototype discovery technique is presented in [6] which includes the processes of Prototype deploying and Prototype evolving, to improve the effectiveness of using and updating discovered Models for finding relevant and interesting information.

Proposed System

3.1. Definitions:

(a) Document set:

Let D be a set of documents, it comprises a set of positive and negative documents. For mining we consider the positive documents only.

$D=D^+ \cup D^-$

(b)Paragraph set:

Each document consists of a finite number of paragraphs. It is called as a paragraph set.

$d= \{tp_1, tp_2,\ldots, tp_m\}$

(c)Term set:

$T= \{t_1, t_2,\ldots, t_n\}$ 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 that containing a particular pattern in 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 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 an individual document and data mining methods are applied. Consider the following Table.1 term sequences in text paragraph in a document.

Table.1 Term sequences in paragraph

Text paragraph	Term sequence
tp_1	t_1 t_2
tp_2	t_3 t_4 t_6
tp_3	t_3 t_4 t_5 t_6
tp_4	t_3 t_4 t_5 t_6
tp_5	t_1 t_2 t_6 t_7
tp_6	t_1 t_2 t_6 t_7

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 algorithm 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 t_2 is found in the document and then t_3 is found then the pattern will be t_2, t_3 and later on t_1 is found then it will be appended and the pattern would be t_2 t_3 t_1.

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, we are going to find the efficacious patterns. Usually the interestingness of a particular user will be known to us 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 to us for searching. After removal of wh question words, prepositions and articles in the search string given by the user we may likely to have the root words only. 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 t_1 (i.e the first term in the search string) and then find the patterns that have the immediate following terms with t_2, t_3 or t_4 etc. It means search for the next three terms after the first term if any of the terms

t_2, t_3, t_4 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 (t_2, t_3, t_4))
6. NP =t_1®x {x is in [t_2, t_3, t_4]}
8. EP=EP U NP
9. if pi[1]=t_2 then
10. search next three terms
11. if(term found is in (t_1,t_3,t_4)) then
12. NP =t_1®x {x is in [t_1, t_3, t_4]}
13. EP=EP U NP
14. end

Initially there are no efficacious patterns, this is indicated by EP=∅. Now taking each pattern in the sequential patterns 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 t_2, t_3 or t_4 exists in the pattern, if so then identify it as efficacious pattern, or else if the first term of the pattern is t_2 then consider the next three terms and check for the three terms t1, t_3, t_4, 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.
Instead of the keyword-based concept used in the traditional document representation model, the pattern based model containing frequent sequential patterns (single term or multiple terms) is used to perform the text mining.

Results
The measures used for evaluating experimental results are ***precision* and *recall* rates**. The *precision* is the fraction of retrieved documents that are relevant to the topic, and the *recall* is the fraction of relevant documents that have been retrieved. These two measures are denoted by the following formulas:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where *TP* (true positives) is the number of documents the system correctly identifies as positives; *FP* (false positives) is the number of documents the system falsely identifies as positives; *FN* (false negatives) is the number of relevant documents the system fails to identify.
Ten topics are chosen for evaluation and the figure.1 shows the number of documents of each topic. The number of frequent sequential patterns are obtained when the minimum support δ=1, δ =2 and δ =2 & pruning is plotted.

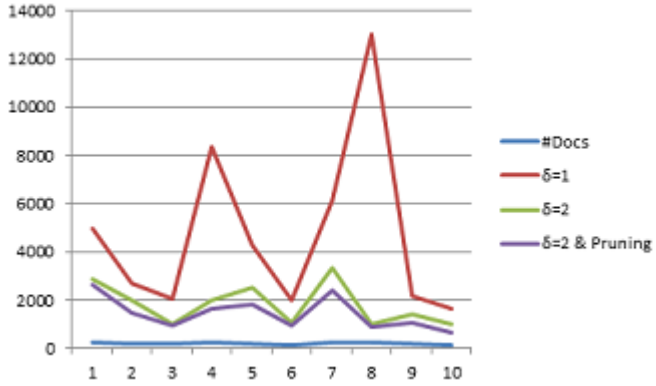

Fig.1 Sequential Pattern graph

The dataset is divided into relevant and irrelevant documents. For relevant documents, the sequential patterns that are correctly identified as positives are 86 and incorrectly identified as negatives are 14. For irrelevant documents the sequential patterns that are incorrectly identified as positives are 19 and correctly identified as negatives are 81.

TABLE.2 Confusion matrix for predicted sequential patterns

Confusion Matrix			
		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	85.39%	14.61%
	Irrelevant	19.29%	80.71%

From the confusion matrix the accuracy, precision and recall rates are calculated. Therefore, from table.3 the accuracy of the sequential pattern algorithm is 83.5% and its precision and recall values are 81.9% and 86% respectively.

TABLE.3 Accuracy, precision and recall

Phases	TP R	FN R	FP R	TN R	Accuracy	Precision	Recall
Seq. Patt.	86	14	19	81	83.5	81.9	86

Conclusion
A novel concept for mining text documents for sequential patterns is presented.The results show that the pattern-based taxonomy model outperforms the keyword-based methods to represent the documents. The results also indicate that removal of meaningless patterns not only reduces the cost of

computation but also improves the effectiveness of the system and also the problem of overfitting is solved.

References

[1] Sharmila, V., I. Vasudevan, And G. Tholkappia Arasu. "Pattern Based Classification For Text Mining Using Fuzzy Similarity Algorithm." *Journal of Theoretical & Applied Information Technology* 63.1 (2014).

[2] Jingle, I., And J. Celin. "Markov Model For Discovering Knowledge In Text Documents." *Journal of Theoretical & Applied Information Technology* 70.3 (2014).

[3] Albathan, Mubarak, Yuefeng Li, and Yue Xu. "Using Extended Random Set to Find Specific Patterns." *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 2. IEEE, 2014.

[4] Kavitha Murugeshan, Neeraj RK. "Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm." *International Journal of Innovative Technology and Exploring Engineering, ISSN*: 2278-3075.

[5] Khade, A. D., et al. "Discover Effective Pattern for Text Mining." *International Journal of Engineering Research and Technology*. Vol. 2. No. 10 (October-2013). ESRSA Publications, 2013.

[6] Archana, K. S., and Aswani Kumar Unnam. "Text Extraction Using Efficient Prototype."

[7] Wu, Sheng-Tang, et al. "Automatic pattern-taxonomy extraction for web mining." *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*. IEEE, 2004.

[8] 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.

[9] D.A.Grossman and O. Frieder, *Information retrieval algorithms and heuristics*, KluwerAcademic publishers, Boston, 1998.