Text Mining Using Pattern Taxonomy Model for Effective Pattern Discovery

Sujatha G S
Department of Computer Science and Engineering
AMC Engineering College
Bangalore, India
gssujatha37@gmail.com

Poonguzhali E
Department of Computer Science and Engineering
AMC Engineering College
Bangalore, India
poonguzhali.e@gmail.com

***Abstract*—Text mining is the discovery of interesting knowledge in text documents. Many data mining techniques have been proposed for mining useful patterns in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. 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 ones. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.**

Keywords—text mining, text classification, information filtering, ptm, pattern deploying, pattern evolving.

I. INTRODUCTION

Data mining is the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. 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 to find particular patterns within a reasonable and acceptable time frame. 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.

Here we focus 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. Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio andprobabilistic models [2], BM25 and support vector machine (SVM) [8] based filtering models.

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like Information. This hypothesis has not fared too well in the history of IR [9], [10]. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include:1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence and 3) there are large numbers of redundant and noisy phrases among them [10].

In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases [7], [13] because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches (or pattern taxonomy models (PTM) [13], [14]) have been proposed, which adopted the concept of closed sequential patterns, and pruned nonclosed patterns. However, the paradox is that people think pattern-based approaches could be a significant alternative, but consequently less significant improvements are made for the effectiveness compared with term-based methods.

There are two fundamental issues regarding the effectiveness of pattern-based approaches: low frequency and misinterpretation. Given a specified topic, a highly frequent pattern (normally a short pattern with large support) is usually a general pattern, or a specific pattern of low frequency. If we decrease the minimum support, a lot of noisy patterns would be discovered. Misinterpretation means the measures used in pattern mining (e.g., "support" and "confidence") turn out to be not suitable in using discovered patterns to answer what users want. The difficult problem hence is how to use discovered patterns to accurately evaluate the weights of useful features (knowledge) in text documents.

In order to solve the above paradox, this paper presents an effective pattern discovery technique, which includes the processes of pattern deploying and pattern evolving, which first calculatesdiscovered specificities of patterns and then evaluates term weights according to the distribution of termsin the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and try 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.

II. RELATED WORK

Many types of text representations have been proposed in the past. A well-known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In [5], the tf*idf weighting scheme is used for text representation in Rocchio classifiers. In addition to tf*idf, the global idf and entropy weighting scheme is proposed and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation approach were given in [1], [4]. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting [10]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. Details of these selection functions were stated in [10].

The choice of a representation is depending on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units [10]. With respect to the representation of the content of documents, some research works have used phrases rather than individual words. The combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase-based text representation for Web document management was also proposed in [12]. In [3], data mining techniques have been used for text analysis by extracting cooccurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms".

Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms, Prefix Span, FP-tree, SPADE, SLPMiner, and GST have been proposed. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemset, cooccurring terms and multiple grams, for building up a representation with these new types of features.

Here the challenging issue is how to effectively deal with the large amount of discovered patterns. For the challenging issue, closed sequential patterns have been used for text mining in [14], which proposed that the concept of closed patterns in text mining was useful and had the potential for improving the performance of text mining. Pattern taxonomy model was also developed in [13] and [14] to improve the effectiveness by effectively using closed patterns in text mining. In addition, a two-stage model that used both term-based methods and pattern based methods was introduced in [6] to significantly improve the performance of information filtering.

Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model [11] was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. This model included three components. The first component analyzed the semantic structure of sentences; the second component constructed a conceptual ontological graph (COG) to describe the sematic structures; and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between nonimportant terms and meaningful terms which describe a sentence meaning. Compared with the above methods, the concept-based model usually relies upon its employed NLP techniques.

III. PROPOSED WORK

Here we consider that all documents are split into paragraphs. So a given document d yields a set of paragraphs PS (d). Let D be a training set of documents, which consists of a set of positive documents, D^+; and a set of negative documents, D^-. Let T = {t1, t2,…,tm} be a set of terms (or keywords) which can be extracted from the set of positive documents, D^+.

In General there are two phases Training and Testing. In training phase the d-patterns in positive documents (D^+) based on a minimum support are found, and evaluates term supports by deploying dpatterns to terms. In Testing Phase to revise term supports using noise negative documents in D based on an experimental coefficient. The incoming documents then can be sorted based on these weights. The proposed approach is used to improve the accuracy of evaluating term weights. Because, the discovered patterns are more specific than whole documents. To avoiding the issues of phrase-based approach to using the pattern-based approach. Pattern mining techniques can be used to find various text patterns. Some important definitions are shown below:

A. Frequent and Closed patterns

A termset X is called frequent pattern if its sup_r (or sup_a) ≥ min_sup, a minimum support. A pattern X (also a termset) is called closed if and only if X = Cls(X).

B. Pattern Taxonomy

Patterns can be structured into a taxonomy by using the is-a (or subset) relation. The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.

C. Closed Sequential Patterns

A sequential pattern X is called frequent pattern if its relative support (or absolute support) ≥ min sup, a minimum support. A frequent Sequential pattern X is called closed if not ∃ any superpattern X_1 of X such that $sup_a(X_1) = sup_a(X)$.

D. Pattern Deploying

The process of interpreting discovered patterns by summarizing them as d-patterns in order to accurately evaluate term weights.

E. Inner Pattern Evolving

The technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem.

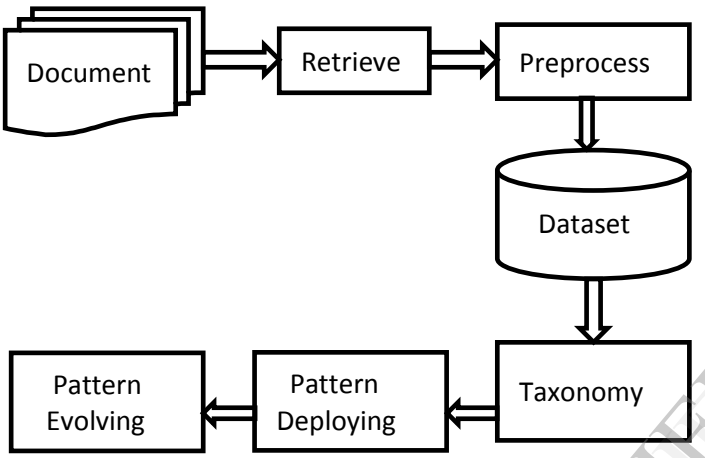

Fig 1: A Frame Work for Proposed System

IV. METHODOLOGY

A. Loading document

Loading document is the process to load the list of all documents. The list of documents is taken from the newsletter collection data set RCV The user is then allowed to retrieve one of the documents from the list of all documents. The retrieved document is then given to the next processcalled preprocessing. Preprocessing can be done to remove the redundant terms.

B. Text preprocessing

Here the retrieved document is subject to preprocessing. Text preprocessing is mainly consists of two types of process to be done on retrieved document.

- Stop word removal
- Stemming

The first process to be carried out is stop word removal, which means Stop words are words which are filtered out prior to, or after, processing of natural language data. The second process to be carried out is text stemming, the process for reducing inflected words to their stem base or root form.

C. Pattern taxonomy process

Here the preprocessed documents are split into paragraphs. So a given document d yields a set of paragraphs PS (d). Let D be a training set of documents, which consists of a set of positive documents, D^+ and a set of negative documents, D^-. Let T= {t1, t2,..., t_m} be a set of terms which can be extracted from the set of positive documents, D^+.Frequent pattern and covering sets are derived from the set of paragraphs and then pattern taxonomy can be build using is-a relation. So from the pattern taxonomy, more semantic information can be extracted.Sample example is as shown below.

TABLE I
A SET OF PARAGRAPHS

Paragraph	Terms
dp_1	{ t_1, t_2}
dp_2	{ t_3, t_4, t_6 }
dp_3	{ t_3, t_4, t_5, t_6 }
dp_4	{ t_3, t_4, t_5, t_6 }
dp_5	{ t_1, t_2, t_6,t_7 }
dp_6	{ t_1, t_2, t_6, t_7 }

TABLE II
FREQUENT PATTERNS AND COVERING SETS

Frequent Pattern	Covering Set
{ t_3, t_4, t_6 }	{ dp_2, dp_3, dp_4 }
{ t_3, t_4 }	{ dp_2, dp_3,dp_4 }
{ t_3, t_6 }	{ dp_2, dp_3, dp_4 }
{ t_4, t_6 }	{ dp_2, dp_3, dp_4 }
{ t_3 }	{ dp_2, dp_3, dp_4 }
{ t_4}	{ dp_2, dp_3, dp_4 }
{ t_1, t_2 }	{ dp_1, dp_5, dp_6 }
{ t_1 }	{ dp_1, dp_5, dp_6 }
{ t_2 }	{ dp_1, dp_5, dp_6 }
{ t_6 }	{ dp_2, dp_3, dp_4, dp_5, dp_6 }

{ t_3, t_4, t_6 }
[2, 3, 4]

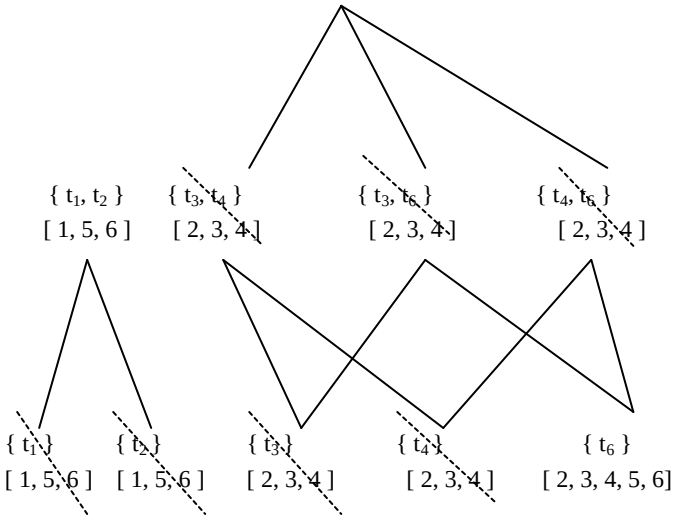

Fig 2: Pattern Taxonomy

D. Pattern deploying

The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated.

E. Pattern evolving

Pattern evolving is mainly used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. Here the noisy pattern named as offender. If partial conflict offender contains in positive documents, the reshuffle process is applied.

V. CONCLUSION

Many data mining techniques have been proposed in the past decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support. We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance.

In this research work, an effective pattern discovery technique has been proposed to overcome the low- frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying which overcome the problem of misinterpretation and pattern evolving which avoids the problem of low-frequency. The proposed technique improves the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.

REFERENCES

[1] K. Aas and L. Eilvil, “K. Aas and L. Eikvil, “Text Categorization: A survey,” Technical Report NR 941, Norwegian Computing Center, 1999.

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

[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), pp. 2-11, 1998.

[4] 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.

[5] X. Li and B. Liu, “Learning to Classify Texts Using Positive and Unlabeled Data,” Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI’03), pp. 587-594, 2003.

[6] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, “A Two-Stage Text Mining Model for Information Filtering,” Proc. ACM 17th Conf. Information and Knowledge Management (CIKM ’08), pp. 1023-1032,2008.

[7] N. Jindal and B. Liu, “Identifying Comparative Sentences in Text Documents,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), pp. 244-251, 2006.

[8] S. Robertson and I. Soboroff, “The Trec 2002 Filtering TrackReport,”TREC,2002,trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.

[9] S. Scott and S. Matwin, “Feature Engineering for Text Classification,”Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 379-388, 1999.

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

[11] S. Shehata, F. Karray, and M. Kamel, “A Concept-Based Model for Enhancing Text Categorization,” Proc. 13th Int’l Conf. Knowledge Discovery and Data Mining (KDD ’07), pp. 629-637, 2007.

[12] R. Sharma and S. Raman, “Phrase-Based Text Representation for Managing the Web Document,” Proc. Int’l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[13]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.

[14] 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.

[15] R.E. Shapire and Y. Singer, “Boostexter: A Boosting-Based System for Text Categorization,” Machine Learning, vol. 39, pp. 135-168, 2000.