Pattern Identification Using Text Mining

Sanil C. Savale
Department of Computer Science,
Gogate Jogalekar college, Ratnagiri
sanil.abc@gmail.com

Anuja. A. Gharpure
Department of Computer Science
Gogate Jogalekar college, Ratnagiri
aagharpure@gmail.com

***Abstract*—Text mining, also referred to as *text data mining*, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, analysis, document, and entity relation modeling (*i.e.*, learning relations between named entities).**

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, mining techniques including link and association analysis, visualization and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Keywords
Clustering, Text mining, pattern mining, pattern evolving, pattern deploying

I. INTRODUCTION

Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as 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. Data mining is therefore an essential step in the process of knowledge discovery in databases. 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.

In this paper, 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. 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. We use the term *text representation* to refer to all issues of representing text content for auto- mated processing. We refer to the set of structures available for representing document content as an *indexing language.* Most current indexing languages represent documents as tuples or vectors of numeric or binary values, with each value corresponding to an indexing term.

II. RELEVANT WORKS.

We use the term *text representation* to refer to all issues of representing text Content for auto-mated processing. We refer to the set of representing. Structures available for representing document content as an *indexing language.* Most current indexing languages represent documents as tuples or vectors of numeric or binary values, with each value corresponding to
an indexing term.

In our previous work, the experimental results showed that Pattern Taxonomy Model (PTM) is a feasible way to apply data mining techniques to the text mining area. However, it is obviously not a desired method for conquering the challenge because of its low capability of dealing with the mined patterns. In our opinion, more robust and effective pattern deploying techniques need to be implemented. Therefore, in this paper we propose two novel pattern deploying algorithms to effectively exploit discovered patterns for the text mining problem.

III. SYNTACTIC PHRASE INDEXING

Syntactic phrase indexing is the use o f syntactic analysis of natural language text to produce multi- word indexing terms. The phrasal term is considered to be assigned to a document only when all its component words appear in the document and have t h e proper syntactic relationship.

1. TERM CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups . It is a main task of exploratory or retrieving data from data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, , image analysis, pattern recognition information retrieval. Work with clustering helps to modify data pre-processing and model parameters until the result achieves the desired properties.

2. TEXT CLASSIFICATION

Text classification is the process of classifying documents into predefined categories based on their content. This paper presents a new algorithm for text classification using data mining that requires fewer documents for training. Text classification is the task of assigning predefined categories to free-text documents. Text classification is used in several fields such as patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on.

IV. DEPLOYING METHOD

In patterns in text mining, we need to interpret discovered patterns by summarizing them as d patterns (see the definition below) in or de r to accurately evaluate term weights (supports). The rational behind this motivation is that d-patterns include more semantic meaning than terms that are selected based on a term based technique (e.g., tf*idf). As a result, a term with a higher tf*idf value could be meaningless if it has not cited by some d patterns (some important parts in documents). The evaluation of term weights (supports) is different to the normal term-based approaches. In the term-based approaches, the evaluation of term weights is based o n the distribution of terms i n documents. In this research, terms a r e weighted according to their a p appearance s in discovered closed patterns.

1. PATTERN TAXONOMY MODEL

As mentioned above, a method is needed to deploy sequential patterns into a feature space. The relation among found patterns can be described as "is-a" pattern taxonomies using PTM. Hence, there are likely many overlaps among these patterns [17]. To represent the overlaps among patterns, we deploy the set of patterns for the document dk o n T, the set of terms, to obtain the following vector:

dk=<(tki , nki),(tk2,nk2),....,(tkm,nkm)>

Where ti in pair (ti , ni) denotes a single term and ni is its support in dk which is the number of patterns that contain ti .

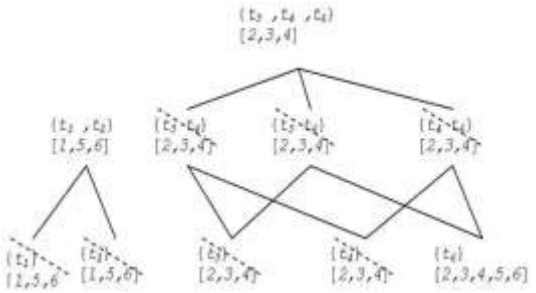

Fig.1.Pattern Taxonomy.

V. D-PATTERN MINING ALGORITHM

To improve the efficiency of the pattern taxonomy mining, an algorithm, SP Mining, wa s proposed in [50] to find al l closed sequential patterns, which used the well-known A priori property in order to reduce the searching space. Algorithm (PTM) shown in fig describes the training process o f finding the set of d-patterns. For every positive document, the SP Mining algorithm is first called in step 4 giving rise to a set of closed sequential patterns SP . The main focus of this paper is the deploying process, which consists of the d-pattern discovery a n d term support evaluation. In Algorithm 1, all discovered patterns in a positive document are composed into a d- pattern giving rise to a set of d-patterns DP in steps 6 to 9. Thereafter, from steps 12 to 19, term supports are calculated
based on the normal forms for all terms in d patterns. To improve the efficiency of the pattern taxonomy mining, an algorithm, SP Mining, was proposed in [50] to find al l closed sequential patterns, which used the well-known A priori property in order to reduce the searching space.

1. INNER PATTERN EVOLUTION

In this section, we discuss how to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. The technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evolution here, because it only changes a pattern's term supports within the pattern. A threshold is usually used to classify documents into relevant or irrelevant categories. Using the d-patterns, the threshold can be defined naturally as follows:

$$Threshold(DP) = \min_{p \in DP} \left(\sum_{(t,w) \in \beta(p)} support(t) \right)$$

There are two types of offenders: 1) a complete conflict offender which is a subset of *nd*; and 2) a partial conflict offender which contains part of terms of *nd*. The basic idea of updating patterns is explained as follows: complete conflict offenders are removed from d-patterns first. For partial conflict offenders, their term supports are reshuffled in order to reduce the effects of noise documents. The task of algorithm Shuffling is to tune the support distribution of terms within a d-pattern. A different strategy is dedicated in this algorithm for each type of offender. As stated in step 2 in the algorithm Shuffling, complete conflict offenders (d-patterns) are removed since all elements within the d-patterns are held by the negative documents indicating that they can be discarded for preventing interference from these possible "noises."

2. BASELINE MODELS

There are two types of Baseline models.
1. Concept based models
2. Term based methods

2.1 CONCEPT BASED MODELS

A new concept-based model was presented in [45] and [46], which analyzed terms on both sentence and document levels. This model used a verb-argument structure which split a sentence into verbs and their arguments. For example, "John hits the ball," where "hits" is a verb, and "John" or "the ball" are the arguments of "hits." Arguments can be further assigned labels such as subjects or objects (or theme). Therefore, a term can be extended and to be either an argument or a verb, and a concept is a labeled term. For a document '*d*' , tf(c)is the number of occurrences of concept c in d; and ctf(c) is called the conceptual term frequency of concept '*c*' in a sentence'*s*', which is the number of occurrences of concept '*c*' in the verb-argument structure of sentences. Given a concept c, its tf and ctf can be normalized as tfweight(c) and ctfweight (c), and its weight can be evaluated as follows:
Weight(c)= t fweight(c) + ct fweight(c).
To have a uniform representation, in this paper, we call a concept as a concept-pattern which is a set of terms. For example, verb "hits" is denoted as {hits} and its argument "the ball" is denoted as {the, ball}

2.2 TERM-BASED METHODS

There are many classic term-based approaches. The Rocchio algorithm [36], which has been widely adopted in information retrieval, can build text representation of a training set using a Centroid. Another well-known term-based model is the BM25 approach, which is basically considered the state-of-the-art baseline in IR .

VI. HYPOTHESES

The major objective of the experiments is to show how the proposed approach can help improving the effectiveness of pattern-based approaches. Hence, to give a comprehensive investigation for the proposed model, our experiments involve
comparing the performance of different pattern based models, concept-based models, and term based models. In the experiments, the proposed model is evaluated in term of the following hypothesis: Hypothesis H1.The proposed model, PTM (IPE), is designed to achieve the high performance for deter- mining relevant information to answer what users want.
The model would be better than other pattern based models, concept-based models, and state-of-the-art term based models in the effectiveness.

. Hypothesis H2.The proposed deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approach is not only promising for pattern-based approaches, but also significant for the concept- based model.

1. EXPERIMENTAL DATA SET

The most popular used data set currently is RCV1, which includes 806,791 news articles for the period between 20 August 1996 and 19 August 1997. These documents were formatted by using a structured XML schema. TREC filtering track has developed and provided two groups of topics (100
in total) for RCV1 [37]. The first group includes 50 topics that were composed by human assessors and the second group also includes 50 topics that were constructed artificially from intersections topics. Each topic divided documents into two parts: the training set and the testing set. The training set has a total amount of 5,127 articles and the testing set contains 37,556 articles. Documents in both sets are assigned either positive or negative, where "positive" means the document is relevant to the assigned topic; otherwise "negative" will be shown. All experimental models use "title" and "text" of XML documents only. The content in "title" is viewed as a paragraph as the one in "text" which consists of paragraphs. For dimensionality reduction, stop word removal is applied and the Porter algorithm [33] is selected for suffix stripping. Terms with term frequency equaling to one are discarded.

2. MEASURES

Several standard measures based on precision and recall are used. 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. The precision of first K returned documents top-K is also adopted in this paper. The value of K we use in the experiments is 20. In addition, the breakeven point (b=p) is used to provide another measurement for performance evaluation. It indicates the point
where the value of precision equals to the value of recall for a topic. The higher the figure of b=p, the more effective the system is. The b=p measure has been frequently used in common information retrieval evaluations.

3. EXPERIMENTAL RESULTS

This section presents the results for the evaluation of the proposed approach PTM (IPE), inner pattern evolving in the pattern taxonomy model. The results of overall comparisons are presented in Table, and the summarized results are described in Fig . We list the result obtained based only on the
first 50 TREC topics in Table since not all methods can complete all tasks in the last 50 TREC topics. As aforementioned, item set based data mining methods struggle in some topics as too many candidates are generated to be processed. In addition, results obtained based on the first 50
TREC topics are more practical and reliable since the judgment for these topics is manually made by domain experts, whereas the judgment for the last 50 TREC topics is created based on the metadata tagged in each document.

The most important information revealed in this table is that our proposed PTM (IPE) outperforms not only the pattern mining-based methods, but also the term-based methods

including the state-of-the-art methods BM25 and SVM. PTM (IPE) also outperforms CBM Pattern Matching and CBM in the five measures. CBM outperforms all other models for the first 50 topics

For the time complexity in the testing phase, all models take O (/t/ * /d/) all incoming documents d. In our experiments, all models used 702 terms for each topic in average. Therefore, there is no significant difference between these models on time complexity in the testing phase.

VII. CONCLUSION

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent item set 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 (i.e., the low frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the in effective 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 and pattern evolving, to refine the discovered patterns in text documents.

REFERENCES

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, a. Norwegian Computing Center, 1999.
[2] R. Agrawal and R. Srikant, "Fast Algorithms for a. Mining Association Rules in Large Databases," Proc. b. 20th Int'l Conf. Very Large Data Bases (VLDB '94), c. pp. 478-499, 1994.
[3] H. Ahonen , O. Heinonen , M. Klemettinen , and a. 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, "Text Categorization with Support Vector Machines:Learning with Many Relevant Features," Proc. European Conf. Machine Learning a. (ICML '98),, pp. 137-142, 1998.
[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and a. C. Gentile, "Kernel Methods for Document b. Filtering,"TREC,trec.nist.gov/pubs/trec11/papers/kermit.ps.gz, 2002.
[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. a. Renders, "Word- Sequence Kernels," J. Machine b. Learning Research, vol. 3, pp. 1059- 1082, 2003.
[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, a. "Statistical Phrases in Automated Text Categorization, b. "Technical Report IEI-B4-07-2000, Instituto di c. Elaborazione dell' Informazione, 297, 1995.
[8] J. Han and K.C.-C. Chang, "Data Mining for Web a. Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[9] C. Cortes and V. Vapnik, "Support-Vecto a. Networks," Machine Learning, vol. 20, no. 3, b. pp. 273-297, 1995.
[10] S.T. Dumais, "Improving the Retrieval of a. Information from External Sources," Behavior b. Research Methods, Instruments, and c. Computers, vol. 23, no. 2, pp. 229-236, 1991