A Novel Approach for Text Mining and An Efficient, Effective Indexing

Ravi Chinapaga
Department of Computer Science and Engineering
TKR Colege f Engineering and Technology, Hyderabad, Telangana, India

M Bal Raju
Department of Computer Science and Engineering
Krishna Murthy Institute Technology & Engineering, Hyderabad, India.

N Subhash Chandra
Department of Computer Science and Engineering
Holy Mary Institute Technology & Science, Hyderabad, India

Abstract- — A lot of mining techniques are proposed for the purpose of mining. The new way for mining in text documents is by using patterns. Anyhow, patterns in data mining have become an open issue for the effectiveness, in the text mining domain mainly. Term-based approaches dependable text mining methods most of all suffer from the crisis of synonymy and polysemy. From the long period, users have often placed the pattern based approaches hypothesis that must and should act better than the ones which are term-based, but lot of researches won't cooperate with this hypothesis. This paper explores an incredible pattern discovery technique which consist the working of pattern placing and also pattern evolving, to increase efficiency of utilizing and patterns updating which was newly discovered for searching relevant and possessive data. The research on TREC contents and RCV1 information collection explains that our proposed system achieves solution effectively. This is useful for clustering large number of web-documents, for clustering large number of Text documents, for clustering large number of research articles.

Keywords – Mining, Pattern, clustering, web documents

I. INTRODUCTION

Due to the rapid growth of digital data made available in recent years, discovery knowledge and information mining have noticed with a huge deal of attention with an imminent necessity for modifying like information into purpose full data and knowledge. Many applications, like analysis of market and management of business, can improve by the usage of the data and knowledge retrieved from a huge amount of information. Discovery of Knowledge can be showed as the way of working nontrivial extraction of data from huge databases, info that is implicitly explored in the data, unknown previously and potentially utilizable for people. Mining of data is by then a perfect step in the way of knowledge discovery process in databases. In the past decade, A lot of mining techniques are proposed for the purpose of mining. All those were consists of sequential pattern mining, frequent itemset mining association rule 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. In this paper, we focus on to solve the misinterpretation problem effective pattern discovery technique is designed. 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. 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, termbased 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 users want.

In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches have been proposed, which adopted the concept of closed sequential patterns, and pruned unenclosed patterns.

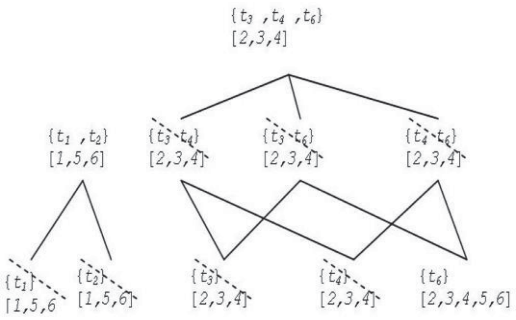

Fig 1: Pattern taxonomy

These pattern mining-based approaches have shown certain extent improvements on the effectiveness. 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 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. Over the years, IR has developed many mature techniques which demonstrated that terms were important features in text documents. However, many terms with larger weights (are general terms because they can be frequently used in both relevant and irrelevant information. For example, term “LIB” may have larger weight than “JDK” in a certain of data collection; but we believe that term “JDK” is more specific than term “LIB” for describing “Java Programming Language”; and term “LIB” is more general than term “JDK” because term “LIB” is also frequently used in C and C++. Therefore, it is not adequate for evaluating the weights of the terms based on their distributions in documents for a given topic, although this evaluating method has been frequently used in developing IR models.

II. PROBLEM STATEMENT

Most existing text mining methods adopted term-based approaches. They all suffer from the problems of polysemy and synonymy. Phrase-based approaches could perform better than the term based ones, as phrases may carry more “semantics” like information.

They have low frequency of occurrence; there are large numbers of redundant and noisy phrases among them. Our proposed system is used to solve the misinterpretation problem effective pattern discovery technique is designed. 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..

1. Paragraph Splitting

The documents contain paragraphs.

The paragraphs are in the unstructured manner.

The module converts paragraphs to structured paragraphs with index.

Example: If a document

d1= {p1, p2, p3,.pn}

The process is applied on the existing document corpus

Word or Term Indexing

Each paragraph of a document is made up with different words or Terms.

Example: S1={t1,t2,t3…….tn}

The module splits all the indexed paragraphs by words.

Unique Words Identification

All the terms in the document corpus are collected by the previous module.

The same words may appear in multiple documents or in the same document.

The module identifies unique words among all the documents and provides index for each term or word as t1, t2, t3, t4 etc,.

4. Paragraph Representation with Unique Terms

Each paragraph can be represented with a sequence of words or terms

5. Frequent Term Sequences.

SP Mining and D-Pattern Mining Algorithms

- D-Pattern Mining algorithm is a kind of Apriori with Frequent Closed Set Approach.
- For every positive document or pattern identified document SPMining Algorithm is called.
- The algorithm prepares all the frequent pattern sequences that occur frequently in the existing corpus.

Algorithm: Pattern taxonomy

input : positive documents D^+; minimum support, min_sup.
output: d-patterns DP, and supports of terms.

```
DP = ∅;
foreach document d ∈ D+ do
    let PS(d) be the set of paragraphs in d;
    SP = SPMining(PS(d), min_sup);
    d̂ = ∅;
    foreach pattern p_i ∈ SP do
        p = {(t, 1) | t ∈ p_i};
        d̂ = d̂ ⊕ p;
    end
    DP = DP ∪ {d̂};
end
T = {t | (t, f) ∈ p, p ∈ DP};
foreach term t ∈ T do
    support(t) = 0;
end
foreach d-pattern p ∈ DP do
    foreach (t, w) ∈ β(p) do
        support(t) = support(t) + w;
    end
end
```

$$\widehat{d_1} = \{(carbon, 2), (emiss, 1), (air, 1), (pollut, 1)\},$$
$$\widehat{d_2} = \{(greenhous, 1), (global, 2), (emiss, 1)\},$$
$$\widehat{d_3} = \{(greenhous, 1), (global, 1), (emiss, 1)\},$$
$$\widehat{d_4} = \{(carbon, 1), (air, 2), (antarct, 1)\},$$
$$\widehat{d_5} = \{(emiss, 1), (global, 1), (pollut, 1)\}.$$

6. Inverse Index Table

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

Let DP be a set of d-patterns in D+, and p ϵDP be a d-pattern. We call p(t) the absolute support of term t, which is the number of patterns that contain t in the corresponding patterns taxonomies.

In order to effectively deploy patterns in different taxonomies from the different positive documents, d-patterns will be normalized using the following assignment sentence:

$$p(t) \longleftarrow p(t) \times \frac{1}{\sum_{t \in T} p(t)}.$$

III. 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 the tf*if weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation approach. 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. 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. The choice of a representation depended on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units. With respect to the representation of the content of documents, some research works have used phrases rather than individual words. In, 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, data mining techniques have been used for text analysis by extracting co occurring 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" as mentioned. Reuters text collection is used to evaluate the proposed approach. Term stemming and stop word removal techniques are used in the prior stage of text pre-processing. Several common measures are then applied for performance evaluation and our results are compared with the state-of-art approaches in data mining, concept-based, and term-based methods. The evaluation of the proposed approach PTM (IPE), inner pattern evolving in the pattern taxonomy model. As aforementioned, itemset-based data mining methods struggle in some topics as too many candidates are generated to be processed. 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|) for 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.

IV.CONCLUSION

A lot mining techniques are proposed for the purpose of mining. All those were consists of sequential pattern mining, frequent itemset mining association rule mining, maximum pattern mining, and closed pattern mining. The new way for mining in text documents is by using patterns (discovered knowledge). The cause is that few long patterns which are useful with max specificity but draw back in low-frequency. We resist that not each and every frequent patterns which are shorts are usable. Cause of patterns misinterpretations of derived from mining techniques gives to the performance which was not perfect. In these experiments, a perfect pattern discovery technique has been introduced to get over the low-frequency and misinterpretation crisis of information mining. The explored technique utilizes couple of processes, pattern evolving and deploying to filter the discovered patterns in text sheets. The researches output explores that the proposed technique processes not only various effective mining-based methods of data or information and the model based on concept, and also even term-based state-of-the-art models, like BM25 and SVM-based models. Proposed technique is efficient in doing or processing such things like identifying Frequent Patterns from text documents takes much time, because the documents have high dimensionality. Each word is an attribute in the document. Frequent Closed Patterns reduces number of iterations while finding frequent patterns. Improves the clustering process by implementing Frequent Closed Set Mining approaches called D-Pattern Algorithm.

REFERENCES

[1] Effective Pattern Discovery for Text Mining Ning Zhong, Yuefeng Li, and Sheng-Tang Wu.

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

[3] 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.

[4] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

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

[6] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile,

[7] "Kernel Methods for Document Filtering," TREC, trec.nist.gov/

[8] pubs/trec11/papers/kermit.ps.gz, 2002.

[9] N. Cancedda, E. Gaussier, C. Goutte, and J.-M.

[10] Renders, "Word- Sequence Kernels," J. Machine

[11] Learning Research, vol. 3, pp. 1059- 1082, 2003.

[12] M.F. Caropreso, S. Matwin, and F. Sebastiani,

[13] "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[14] C. Cortes and V. Vapnik, "Support-Vector

[15] Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[16] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior

[17] Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

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

[19] 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.

[20] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp.

[21] 4-9, 2003.

[22] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l

[23] ACM SIGIR Conf. Research and Development in

[24] Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[25] Categorization," Proc. 14th Int'l Conf. Machi