Novel Approach to Discover Effective Patterns For Text Mining

Rujuta Taware
ME-II Computer Engineering,
JSPMS's BSIOTR (W), Wagholi,
Pune, India.

Prof. Sanchika A. Bajpai
Department of Computer Engineering,
JSPMS's BSIOTR (W), Wagholi,
Pune, India.

Abstract- **Many data processing techniques are proposed for mining helpful patterns in text documents. However, how to effectively use and update those discovered patterns remains an open analysis issue, particularly within the domain of text mining. The majority existing text mining ways adopted term-based approaches; all of them suffer from the issues of ambiguity and synonymousness. This paper presents innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of victimization and change discovered patterns for locating relevant and attention-grabbing data.**

Keywords: Text Mining, Pattern Mining, Pattern Taxonomy, Pattern Evolution.

I INTRODUCTION

A. Overview

Due to the rise of knowledge created obtainable in recent years, information discovery and data processing have attracted a good deal of attention with associate close need for turning such knowledge into helpful data and knowledge. Several applications, like market research and business management, will profit by the employment of the information and information extracted from an outsized quantity of data. Information discovery will be viewed because the method of nontrivial extraction of data from massive databases, information that's implicitly conferred within the knowledge, previously unknown and probably helpful for users. Data mining is so a vital step within the method of knowledge discovery in databases. Many types of data mining techniques are used association rule mining, sequential pattern mining and closed sequential pattern etc.

B .Problem Definition

Data mining is the process of retrieving interesting knowledge from database. In the proposed system, the discovery of patterns will be done efficient through the pattern evolution and pattern deploying. The system will not only find the useful patterns but also efficiently use and update them to find relevant and interesting information. The problem of low frequency and misinterpretation will be solved. The system is supposed to develop the knowledge discovery model which will efficiently use and update the patterns.

II LITERATURE SURVEY

Study on Term based approach

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 users want.

1. Multi-Tier Granule Mining for Representations of Multidimensional Association Rules
Author: Y. Li, W. Yang, and Y. Xu.

Year: Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.
Description: Machine-learning techniques play the important roles for information Filtering. The main objective of machine-learning is to obtain users profiles. To decrease the burden of on-line learning, it is important to seek suitable structures to represent user information needs. This paper proposes a model for information filtering on the Web. The user information need is described into two levels in this model: profiles on category level, and Boolean queries on document level. To efficiently estimate the relevance between the user information need and documents, the user information need is treated as a rough set on the space of documents. The rough set decision theory is used to classify the new documents according to the user information need. In return for this, the new documents are divided into three parts: positive region, boundary region, and negative region.

Advantages:

1. A new approach for research into IF models is presented in this paper. The new approach allows users to describe their information needs on user concept spaces rather than on the space of documents. Therefore, the task of IF models is to build the relationships between user concept spaces and the spaces of documents. A rough set based IF model has been proposed in this paper, which views the user information need as an approximate concept (a rough set) over the space of documents.

2. The rough set based IF model has been used in the JobAgent to solve the information overload problem. In JobAgent, the user describes the information need in relation to a certain concept structure. The connections between the user information need and the different Web pages can be built by a rough set based IF model.

3. A document clustering algorithm which can decide the category for a given document has been presented in JobAgent . This algorithm uses some centroids (more than one) to represent a category, and introduces the concept of difference degrees to group documents.

4. Different from the probabilistic IR models, the rough set based IR model puts the terms in different levels (based on the user concept) rather than defining the probabilistic relation between terms.

Disadvantages**:**

1. Failed to find a suitable model to explain the term's probabilities by using the user concept.

Study on Phrase Based Approach

It was assumed that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. 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.
3) There are large numbers of redundant and noisy phrases among them .

1. Feature Selection and Feature Extraction for Text Categorization

Author: D.D. Lewis.

Year: Proc. Workshop Speech and Natural Language, pp 212-217, 1992.

Description: The effect of selecting varying numbers and kinds of features for use in predicting category membership was investigated on the Reuters and MUC-3 text categorization data sets. Good categorization performance was achieved using a statistical classifier and a proportional assignment strategy. The optimal feature set size for word-based indexing was found to be surprisingly low (10 to 15 features) despite the large training sets. The extraction of new text features by syntactic analysis and feature clustering was investigated on the Reuters data set. Syntactic indexing phrases, clusters of these phrases, and clusters of words were all found to provide less effective representations than individual words. The indexing language used to represent texts influences how easily and effectively a text categorization system can be built, whether the system is built by human engineering, statistical training, or a combination of the two. The simplest indexing languages are formed by treating each word as a feature. However, words have properties, such as synonymy and polysemy that make them a less than ideal indexing language. These have motivated attempts to use more complex feature extraction methods in text retrieval and text categorization tasks.

Advantages:

1] Feature extraction and selection is done.

2] Statistical classifier trained on manually categorized documents to achieve quite effective performance in assigning multiple, overlapping categories to documents is proposed.

3] It is shown that via studying text categorization effectiveness, a variety of properties of indexing languages that are difficult or impossible to measure directly in text retrieval experiments, such as effects of feature set size and performance of phrasal representations in isolation from word-based representations.

Disadvantages:

1] For text categorization, in particular the ineffectiveness of term clustering with coarse-grained meta features, are likely to hold for text retrieval as well.

2. Feature Engineering For Text Classsification

Author: S. Scott and S. Matwin.

Year: Proc. 16th Int'l Conf. Machine Learning (ICML'99), pp. 379- 388, 1999.

Description: Most research in text classification has used the "bag of words" representation of text. This method examines some alternative ways to represent text based on syntactic and semantic relationships between words (phrases, synonyms and hypernyms). It describes the new representations and tries to justify our suspicions that they could have improved the performance of a rule-based learner. The representations are evaluated using the RIPPER rule-based learner on the Reuters-21578 and DigiTrad test corpora, but on their own the new representations are not found to produce a significant performance improvement. Finally, it tries combining classifiers based on different representations using a majority voting technique. This step does produce some performance improvement on both test collections. In general, this work supports the emerging consensus in the information retrieval community that more sophisticated Natural Language Processing techniques need to be developed before better text representations can be produced.

Advantages:
1] It provides alternative representations could serve as the basis for combining classifiers to produce better results.

Disadvantages:
1] It concerns over the use of micro-averaged breakeven point in the information retrieval literature.

3. Machine Learning in Automated Text Categorization

Author: F. Sebastiani

Year: ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

Description: The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories.

Advantages:

1] The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert manpower, and straight forward portability to different domains.

Disadvantages:

1] Multimedia indexing problem are not solved in a satisfactory way, the general methodology for text also apply to automated multimedia categorization.

Study on Pattern Based Approach

1. Identifying Comparative Sentences in Text Documents

Author: N. Jindal and B. Liu

Year: Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

Description : This paper studies the problem of identifying comparative sentences in text documents. The problem is related to but quite different from sentiment/opinion sentence identification or classification. Sentiment classification studies the problem of classifying a document or a sentence based on the subjective opinion of the author. An important application area of sentiment/opinion identification is business intelligence as a product manufacturer always wants to know consumers' opinions on its products. Comparisons on the other hand can be subjective or objective. Furthermore, a comparison is not concerned with an object in isolation. Instead, it compares the object with others. Identifying comparative sentences is also useful in practice because direct comparisons are perhaps one of the most convincing ways of evaluation, which may even be more important than opinions on each individual object.
Advantages:

1] It studies identifying comparative sentences. Such sentences are useful in many applications, e.g. marketing intelligence, product benchmarking, and e-commerce.

2] It analyzes different types of comparative sentences from both the linguistic point of view and the practical usage point of view, and shows that existing linguistic studies have some limitations.

Disadvantages:

1] It fails to prove both the precision and recall of the proposed technique.

2. Deploying Approaches for Pattern Refinement in Text Mining.

Author: S.-T. Wu, Y. Li, and Y. Xu.

Year: Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006

Description: Text mining is the technique that helps users find useful information from a large amount of digital text documents on the Web or databases. Instead of the keyword-based approach which is typically used in this field, the patternbased model containing frequent sequential patterns is employed to perform the same concept of tasks. However, how to effectively use these discovered patterns is still a big challenge. In this study, it is proposed that two approaches based on the use of pattern deploying strategies. The performance of the pattern deploying algorithms for text mining is investigated on the Reuters dataset RCV1 and the results show that the effectiveness is improved by using our proposed pattern refinement approaches.

Advantages:

1] This method proposes two pattern refinement methods to deploy the discovered patterns into a feature space which is used to represent the concept of documents.

2] This method adopts the mining sequential pattern technique to find semantic patterns from text documents and then deploy these patterns using proposed deploying algorithms.

Disadvantages:

1] Lacks in learning the change of the characteristics of context .

3. Automatic Pattern-Taxonomy Extraction for Web Mining

Author: S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen.

Year: Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004

Description: In this paper, it proposes a model for discovering frequent sequential patterns, phrases, which can be used as profile descriptors of documents. It is indubitable that we can obtain numerous phrases using data mining algorithms. However, it is difficult to use these phrases effectively for answering what users want. Therefore, it presents a pattern taxonomy extraction model which performs the task of extracting descriptive frequent sequential patterns by pruning the meaningless ones. The

model then is extended and tested by applying it to the information filtering system. The results of the experiment show that pattern-based methods outperform the keyword-based methods.

Advantages:

1] Removal of meaningless patterns not only reduces the cost of computation but also improves the effectiveness of the system.

Disadvantages:

1] To improve the accuracy, one major direction is to extract and use the interesting information from negative or unlabeled documents. The weighting scheme of discovered patterns is not optimized.

III Proposed System

1. An effective pattern discovery technique, is discovered .

2 .Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns.

3. Solves Misinterpretation Problem

4. Considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries to reduce their influence for the low-frequency problem.

5. The process of updating ambiguous patterns can be referred as pattern evolution.

6. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.

7. In training phase the d-patterns in positive documents (Dþ) based on a min sup are found, and evaluates term supports by deploying dpatterns to terms

8. In Testing Phase to revise term supports using noise negative documents in D based on an experimental coefficient

9. The incoming documents then can be sorted based on these weights.

Advantages of proposed system:

1. The proposed approach is used to improve the accuracy of evaluating term weights.

2. The discovered patterns are more specific than whole documents.

3. To avoid the issues of phrase-based approach to using the pattern-based approach.

4. Pattern mining techniques can be used to find various text patterns.

System Architecture:

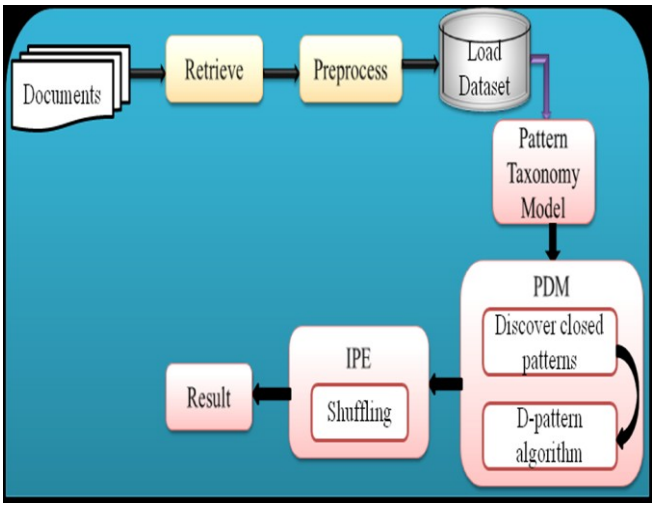

Figure: System Architecture

Figure contains the following blocks:

1. Loading document: In this module, to load the list of all documents. The user to retrieve one of the documents. This document is given to next process. That process is preprocessing.

2. Text Preprocessing: The retrieved document preprocessing is done in module. There are two types of process is done.1) stop words removal 2) text stemming .Stop words are words which are filtered out prior to, or after, processing of natural language data. Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

3. Pattern taxonomy process: In this module, the documents are split into paragraphs. Each paragraph is considered to be each document. In each document, the set of terms are extracted.The terms, which can be extracted from set of positive documents.

4. 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.

5.Pattern evolving :In this module used to identify the noisy patterns in documents.Sometimes, system falsely identified negative document as a positive.So, noise is occurred in positive document.The noised pattern named as offender.If partial conflict offender contains in positive documents, the reshuffle process is applied.

IV Conclusion

There were lots of data mining techniques from past decades but, the updating of discovered pattern effectively was difficult with those techniques because the long pattern with high specificity lacks in support. Inadequate use of patterns that are extracted also causes performance degradation. The main issue regarding the pattern based

approach is low frequency and misinterpretation. In order to enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection This presents research pattern taxonomy model which includes pattern evolving and deploying method helps in the updating of useful pattern efficiently and the two issues can be solved. It helps in finding the useful information to the user. The inner pattern evolution outperforms the pattern deploying method.

REFERENCES

[1]Y. Li, W. Yang, and Y. Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 953-958, 2006.

[2] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

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

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

[5] 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.

[6] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. DataMining (ICDM '06), pp. 1157-1161, 2006

[7]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.

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

[9] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'lConf. Web Intelligence (WI '04), pp. 242-248, 2004.

[10] S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" In iiWAS'05, pages 1013–1022, 2005.

[11] 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.

