Survey Paper on Pattern Discovery Text Mining for Document Classification

Gaikwad Varsha R. 
IT Department 
Dr.D.Y.Patil Institute of Engineering and Technology, Ambi, Pune

Patil Harshada R. 
IT Department 
Dr.D.Y.Patil Institute of Engineering and Technology, Ambi, Pune

Vibha B. Lahane 
IT Department 
Dr.D.Y.Patil Institute of Engineering and Technology, Ambi, Pune

ABSTRACT

In text documents data mining techniques have been discovered for mining useful patterns. But there are some questions, how to properly use and update discovered patterns is still an open research issue, specifically in the text mining. Therefore most existing text mining methods have used term-based approaches but they all suffer from the problems of polysemy (multiple meaning word) and synonymy(same meaning word). This is the literature survey paper with proposed system develops innovative and successful pattern deploying technique which contains the processes of pattern taxonomy, pattern deploying and gradually developing pattern, to improve the effectiveness of using and updating researched patterns for finding applicable and interesting data with effective patterns as per the users requirements. In this paper user is also getting the meaningful information without wrong meaning problem.

General Terms
Pattern Taxonomy, D-matrix

Keywords
Text mining, information filtering, Pattern Mining, Data Mining, Text Classification, pattern evolving

1. INTRODUCTION

In the last few years, many data mining techniques have been presented in order to perform different data separating tasks. These cover ARM (association rule mining), CLOSET (Closed Often Item set), SPM (Linear pattern Mining), MPM (maximum pattern mining), and CPM (Closed Pattern Mining). There is fast growth of digital data made available in last few years; data discovery and data mining have work together to get extraction of meaningful data into useful information and knowledge. Data mining is therefore an effectual step in the process of knowledge discovery in databases. Text mining is used to finding applicable & interesting information from huge database. Text mining is to utilize information contained in textual documents in different ways including discovery of patterns, association among entities, etc. With this survey paper we focus on the literature survey of a pattern based data discovery model. The benefits of term based methods contain efficient computational performance as well as theories for term weighting.

2. LITERATURE SURVEY


In information retrieval, Okapi BM25 is a ranking method used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic regains framework developed in the 1970s and 1980s by Karen Spärck Jones, Stephen Robertson and others. BM25 is a bunch of words recuperation method that ranks a set of documents based on the query terms appearing in each document nevertheless of the connection in between the query terms within a document. It is not a single method, but actually a whole family of scoring methods, with little different elements and attributes. There is a large group of documents, say, [13] at web level; the Boolean model either regains too many documents or very few documents. This is because: users usually do not form complex queries. Either they use very few (often a single) term fetching a tremendously large list of unordered documents. Else, they use a large set of terms joined by AND. This fetches very few documents. Bhattacharyya, 2008).

2.2 TFIDF (1980’s)

TF-idf, [5] short for term rate of occurrence-inverse document rate of occurrence, is a numerical statistic that is intended to reflect how important a word is to a document in a gathering or collection. It is also used as a weighting factor in data regain and text mining. The TF-IDF value growsrationally to the number of times a word appears in the document, but is equalized by the rate [8] of occurrence of the word in the collection, which helps to control for the actual thing that some words are commonly more general than others.

2.3 CLOSET (1993)

There is an interesting substitute,[3] proposed by Pasquier et al. instead of mining the complete set of frequent item sets and their associations, association mining only needs to end often closed item sets and their equivalent rules. A major intimation is that mining often closed item sets has the same power as mining the complete set of frequent items, but it will considerably reduce unwanted rules to be produced and enlarges both productivity and success of mining.

2.4 Support Vector Machine (1995)

The support vector machine (SVM) is a training algorithm for learning classification and regression rules from data,[6] for example the SVM can be used to learn polynomial, (RBF) radial basis method and (MLP) multi-layer perceptron classifiers. SVMs were first suggested by Vapnik in the 1990s for classification and have recently become an area of intense research owing to [14] developments in the techniques and theory coupled with extensions to regression and solidity estimation. SVMs came about from analytical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as a median step. These are based on the configurationally risk minimization principle, closely related to regularization theory. This principle incorporates capacity control to prevent over fitting and thus is a partial solution to the bias-variance trade-off dilemma. A common disadvantage of non-parametric techniques such as
SVMs is the lack of placidity of results.[16] Support vector machines (SVMs) cannot represent the score of all companies as a simple parametric method of the economic ratios, since its measurements may be on top. It is neither a linear conjunction of single economic ratios nor has it another simple methodological form. The weights of the economic ratios are not constant. Thus the marginal benefaction of each economic ratio to the score is variable.

2.5 Sequential Pattern Mining (2004)

The linear pattern mining problem was first introduced by Agrawal and Srikant.[11] Given a set of series, where each series consists of a list of elements and each element consists of a set of items, and [2] given a [15] user-specified mini support approach, linear pattern mining is to find all often subsequences, i.e., the subsequences whose phenomenon frequency (i.e. rate of occurrence) in the set of series is no less than mini support.

2.6 CBM (2010)

CBM is based on Ohlsson's theory of learning from performance errors. CBM focuses on faulty data, realizing that it is not enough to describe what the student knows is right. The basic presumption is that analytic information is not hidden in the order of student's activity, but in the problem state the student came up at. [9] This presumption is carried by the actual thing that there can be an incorrect solution of a problem that negotiates a problem state, which disobeys radical ideas, or notions of the domain. There is a demerit in calculation of shift using three different processes and taking high of these increases elevated in searching.

3. PROPOSED METHODOLOGY

In the field of text mining, pattern mining techniques can be used to find various text patterns, such as linear patterns, often item sets, co-existing terms and multiple grams, for putting up a description with these new types of specifications. However, the demanding issue is how to productively deal with the large amount of uncovered patterns. For the demanding issue, closed linear patterns have been used for text mining in, which initiated that the notion of closed patterns in text mining was useful and had the probable for raising the presentation of text mining. [11] Pattern taxonomy structure was also developed in ZHONG ET AL.: EFFECTIVE PATTERN DISCOVERY FOR TEXT MINING and to upgrade the success by effectually using closed patterns in text mining. In addition, a two-stage structure that used both term-based process and pattern based process was established in to importantly improve the presentation of information refining. 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 scuffling for dealing with uncertainties in human languages. Recently, a new concept based structure was presented to fill the gap between NLP and text mining, which inspected terms on the sentence and document positions. This structure contained three elements. The first element analyzed the linguistic structure of sentences; the second element constructed Theoretical ontological graph (TOG) to describe the linguistic structures; and the last element separated top concepts based on the first two elements to build specification vectors using the standard vector space structure. The benefit of the concept-based structure is that it can productively differentiate between unimportant terms and meaningful terms. Equated with the above process, the concept-based structure usually depends upon its employed NLP techniques.

So Here, We are adopting the new structures established in ZHONG ET AL.: EFFECTIVE PATTERN DISCOVERY FOR TEXT MINING

- Pattern Taxonomy Structure:
  - Pattern Taxonomy
  - Closed Linear Patterns
- Pattern Deploying Process:
  - D-Pattern matrix
- Inner Pattern Evaluation:

4. CONCLUSION

Many data mining techniques have been initiated in the last decade. These techniques carry (ASM) association rule mining, (CLOSET) closed frequent item set mining, maximum pattern mining, (SPM) sequential pattern mining and closed pattern mining. However, using these uncovered data (or patterns) in the field of text mining is hard to implement and not as much effective. This is because some useful long patterns with high specificity minimum support (i.e., the low-rate of occurrence problem). We argue that not all recurrent short patterns are useful. Hence, misapprehension of patterns obtained from data mining techniques lead to the unsuccessful presentation. In this research work, an effectual pattern discovery technique has been established to overcome the low rate of occurrence and misapprehension problems for text mining. This proposed technique uses two processes, pattern evolving and pattern deploying, to refine the uncovered patterns in text documents. The exploratory results show that the proposed structure out performs not only other pure data mining-based process and the concept based structure, but also term-based state-of-the-art structures, such as BM25 and SVM-based structures.

5. REFERENCES

[2] Jian Pei, Member, IEEE Computer Society, Jiawei Han, Senior Member, IEEE, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Member, IEEE Computer Society, and Mei-Chun Hsu 2004 “Mining sequential Patterns by Pattern-Growth: The Prefix Span Approach”
[3] Jian Pei, Jiawei Han, and Runyong Mao 2010 “CLOSET: An Efficient Algorithm for Mining Often Closed Item sets”


[16] Laura Auria Rouslan A. Moro August 2008 “Support Vector Machines (SVM) as a Technique for Solvency Analysis”