EFFECTIVE OPTIMAL SEARCHING MECHANISM USING TEXT MINING

J.E Nalavade,Ruchi Priya,Priya Koul,Riya and Manish Anand ,University of pune,SIT Lonavala,Computer Science,

Abstract—For many data mining techniques have been discovered early which was use in mining useful text pattern in documents. However,there is still a research issue,how to effectively discovered patterns in a text, especially in the domain of text mining.Most of the existing text mining methods adopted in term-based approaches,but all these techniques suffer from the problems of polysemy and synonymy. Over theyears, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.In this paper presents an effective an optimal searching technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of optimal searching patterns inorder to finding relevant and interesting information from large database also introduce a class of simple combinatorial patterns over the phrases, called proximity phrase association patterns, and consider the problem of fndingthe patterns that optimizes a given statistical measure in a large collection of unstructuredtexts. For this class of patterns, we develop fast and offset based text mining algorithms which will use for optimal search of text document.

Keywords- Text mining, Pattern mining, Text classification, Pattern evolving, Information filtering, K optimal search.

1. INTRODUCTION

For many NLP(natural language processing) tasks,These are mostly useful resources which have very high precision entries but have some important limitations. when they used in real-world NLP tasks due to their limited coverage and prescriptive nature.Verbs are the primary vehicle for describing events and expressing relations between entities. Hence, verb semantics could help in many natural language processing (NLP) tasks that deal with events or relations between entities. For this tasks it will require canonicalization of natural language statements or derivation of plausible inferences from such statements, a particularly valuable resource is one which (i) relates verbs to one another and (ii) provides broad coverage of the verbs in the target language.So we develop fast and optimal text mining algorithms based on techniques from computational offer and base recursive techniques. Then, we made experiments on large collections of documents and on Web pages to evaluate the proposed method.In this techniques we include association rule mining, sequential pattern mining, maximum pattern mining, and closed pattern mining, frequent itemset mining. The purpose of developing efficient and optimal mining algorithms to find particular patterns within a reasonable and acceptable time frame.With a large number of patterns generated by using the approachs of data mining, however to optimize the use of this patterns in real life.There are two fundamental issues regarding the techniques which alreay introduce in the market are: low frequency and misinterpretation and location.And this 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 the question what the users want. The difficult problem hence is how to use discovered patterns to accurately evaluate the weights of useful features(knowledge) in text documents.

The BM25 and support vector machine (SVM) are based filtering models. The advantages of these termbased methods include efficient computational performance as well as mature theories for the term weighting,which have emerged over the last discovered paper from the IR (information retrive)and machine learning communities.But however, termbased techniques suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy means multiple words having the same meaning.Due to this problem user did not get the answer what they want.After that the discover of phrase-based approaches could perform better than the termbased ones, as phrases may carry more “semantics” like information.But it again discouraging performance of the system the reason behind that: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.

To overcome the disadvantages of phrase-based approaches,pattern mining-based approaches (or pattern taxonomy models (PTM) [1]) have been proposed, which adopted the concept of closed sequential patterns.These pattern mining-based approaches have shown certain extent improvements the of effectiveness of the system. However, there is a contradiction that people thinkthat pattern-based approaches could be a significant alternative,but it consequently less significant where the improvements are made effectively by compared with term-based methods.

There are two fundamental issues regarding the effectiveness of pattern-based approaches: low frequency and misinterpretation and highly frequent pattern (normally

a short pattern with large support) is a general pattern, or a specific the pattern of low frequency. If we reduce the minimum support, a lot of noisy patterns would be discovered but again there is a problem of misinterpretation means that the measures used in pattern mining (e.g., "support"and "confidence") turn out to be not suitable in using discovered patterns to answer the users question.it is still a difficult problem, hence how to use discovered an effective approach which will be over come the previous discovered pattern.

In this paper presents an effective optimal searching pattern techniques, which will first calculates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution of documents for solving the problem of misinterpretation. The approach can improve the accuracy of evaluating term weights because to discovered patterns which are more specific than the whole documents.The Reuters Corpus Volume 1 (RCV1) and Text Retrieval Conference (TREC) to filtering topics, in order to evaluate the proposed technique.In this the results show that the proposed technique outperforms up-to-date of data mining-based methods,and concept-based models and the state-of-the-art term based methods.

On the other hand the rest part of this paper are present,In Section 2 discusses related work. Section 3 some discussion about Proximity phrase association patterns regarded as a generalization of the association rules in transaction of databases and about the actual of the discussion text mining [1] such that (i) each item is a phrase of arbitrary length, (ii) items are ordered, and (iii) a proximity constraintis introduced.Section .Section 4 provides some definitions and advantage ,disadvantage of previous discovered pattern5 the proposed system of and how it will effective from previous discovered pattern and how offset algorithm use.Section 6describe its application.Section 7.presents experimental setting and results for evaluating the proposed approach. Finally,Section 8 given its reference paper.

2. RELATED WORK

There are many types of text mining techniques have been proposed in the past. And the well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. The tf*idf weight scheme is used for text representation . In the addition of TFIDF,the global IDF and entropy weighting scheme is to be proposed in [9] and improves performance by an average of 30 percent. Various schemes for representation of the bag of words by weighting approach were given in [1]. The problem of the bag of words, how to select the proper limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid overfitting .And to reduce the number of features,there are 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 statedAnd. the problem of Knowledge Discovery from Text (KDT) [9] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into corpus of text data. KDT, while deeply rooted in NLP,and draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding.

Text mining is similar to data mining techniques, except

The choice of a representation depended on what regards as the meaningful text and the meaningful natural language rules for the combination of these units .With respect to the representation of the content of documents, and some research works have used phrases rather than individual words.The combination of unigram and bigrams was chosen for document indexing in the text categorization of (TC) and evaluated on a variety of feature evaluation functions (FEF). In [3], data mining techniques have been used for text analysis and by extracting co-related terms as a descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed a number of significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned in term-based ontology mining methods also provided some thoughts for text representations. For example, hierarchical clustering was used to determine synonymy and hyponymy relations between keywords. Also, the pattern evolution technique (PTM) was introduced in [1] in order to improve the performance of term-based ontology mining.Pattern mining has been extensively studied in data mining. A variety of efficient algorithms such as Apriori-like algorithms, PrefixSpan [11], [16], FP-tree [15], [14],SPADE[17], SLPMiner [10], 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,optimal searching have been purposed here inorder to make useful and interesting patterns and rules are discovered. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets, co-relating terms and multiple grams, for building up a representation with these new types of features.

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 [12], 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 inand [12] 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 patternbased methods was introduced and improve the performance of information filtering. Natural language processing (NLP) is a modern computational technology that can also help people to understand the meaning of text documents. Very long time,

NLP was struggling for dealing with uncertainties in human languages.But recently, a new concept-based model 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. And compared with the above methods,the concept-based model usually relies upon its employed NLP techniques.

3 . TEXT MINING

Text mining is a process of extracting smart text from a corpus data.mining is a tcchnique,that tries to find out the interesting patterns from large databases. Text mining is also use in various of field.Text mining, also known as discovery of Intelligent Text Analysis,In text Data Mining or Knowledge-Discovery in Text (KDT), which refers generally to the process of extracting interesting and structured information and knowledge from the unstructured text. Text mining is a young inter disciplinary field which will be draws on information retrieval, data mining, machine learning, statistics and computational relating to the language. But most the information (over 80%) are stored as text document, text mining is believed to have a high commercial potential value in the mining technique. Knowledge may be discovered from many sources of information but it is not necessary that all structured text there may be unstructured texts remain the largest readily available source of knowledge.Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases , but text mining process can work with non trival or trival data sets such as emails, full-text documents and Web files etc. The result of a text mining is a much better solution for different field. However,the most or research and development efforts have been centered on data mining techniques and efforts by using structured data. The problem introduced by text mining is : natural language (NPL) was developed for humans to communicate with one another and record and information,and computers are a long way to figure out natural language.If the information extracted from a corpus of documents then it will represents abstract knowledge rather than concrete data, IE itself can be considered as a “discovering” knowledge from text.Text mining process have been used to analyze research publications as well as electronic patient records. It will also used in Biomedical entities such as drug names, proteins information, genes, and diseases can be automatically extracted from stored documents and used to construct a geniual pathways or to provide mapping into existing medical ontologies.Text mining aims to extract useful knowledge from textual data or documents (Hearst, 1999; Chen, 2001). Although text mining is also considered a subfield of data mining, some of the text mining techniques have originated from other disciplines, such as information retrieval,visualization of information, computational linguistics, and information science. Examples of text mining applications include document classification entity extraction ,document clustering, information extraction, and summarization.Knowledge Management, Data Mining and Text Mining Most knowledge management, data mining, and text mining techniques involve learning patterns from existing data or information, and therefore they built the foundation of machine learning and artificial intelligence. Inthe following, we review several major paradigms in machine learning, important evaluation methologies.

Let us consider a example of text mining in order to explain the actual meaning of text mining let us consider a web site like Twitter1 text of DataMining is used as the data to analyze. It starts with extracting text from the Twitter. It extracted text is then transformed to build a document-term matrix. And after that, frequent words and associations are found from the matrix. A word cloud is used to present important words in a documents. And at the end of the words and tweets are clustered to find a groups of words and also groups of tweets data. In this "tweet" and document" will be used interchangably, so that "word" and "term" are extracted. There are three packages used : twitter,tm and wordcloud. Package twitter to provides access to Twitter data, tm provides functions for text mining, and wordcloud visualizes the result with a word cloud from text data.

4. DESIGN STREAMS

4.1 The First step in Search TechnologyIn Text Mining

By Salton’s in the year of 1968 :Here finding text without knowing exactly what user are looking and finding what apparently is not there Text mining is particularly interesting in areas where users have to discover new information. In this case, lets an example, in criminal investigations, legal discovery and due diligence investigations. Such investigations require 100% recall, i.e., users can not afford to miss any relevant information. In contrast, a user searching there keyword, internet for background information using a standard search engine which simply the requires any type of information (a oppose to all information) as long as it is reliable. In a due diligence, a lawyer certainly wants to find all possible liabilities and is not interested in finding only the obviousones.Here documents and search in terms and converted to vectors and compared by using the cosine distance between them: how smaller the cosine distance, how more the search term and the corresponded document. This was an effective method to determine the relevance of a document from the search term. This was called the *vector*

space model, but it is time consumint and not effective method .

4.2 Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections

By Helena Ahonen and Oskari Heinonen Mika Klemettinen A. Inkeri Verkamo:they proposed The frame-work follows the general KDD process, thus containing steps from preprocessing to the utilization of the results.Both pre- and post processing have essential roles in pruning and weighting the results.but it again suffer from the problem of The KDT system and Document Explorer used in mining Reuters news articles.It requires a substantial amount of background knowledge, and is not applicable as such to text analysis in general.

4.3Automatic Text Categorization and its Application for Text Retrieval

By WaiLam , MiguelRuiz and Padmini Srinivasan: they proposed The categorization approach derives from a machine learning paradigm. Known as instance based learning and an advanced document retrieval technique known as retrieval feedback.It also investigates the application of this categorization process to advanced text retrieval.But this syatem suffer the problem of

4.4 A Boosting-based System for Text Categorization

By Robert Eschapire and Yoram Singer:He proposed The use of a machine-learning technique called boosting to the problem of text categorization. The main idea of boosting is to combine many simple and moderately inaccurate categorization rules into a single, highly accurate categorization rule.But the system again suffer from the problem of Text categorization is the problem of classifying text documents into categories or classes. For instance, a typical problem is that of classifying news articles by topic based on their textual content.

4.5 Mining Generalized Association Rules

By Ramakrishnan Srikant and RakeshAgrawal:they proposed the solution to the problem is to replace each transaction with an extended transaction.It contains all the items in the original transaction as well as all the ancestors of each item in the original transaction.But there is a problem of In existing association rules did not consider the presence of taxonomies, and restricted the items in the association rules to the leaf-level items in the taxonomy. This association rule is not very fast.

4.6 Text Classification using String Kernels

By Huma Lodhi, John Shawe-Taylor, Nello Cristianini and Chris Watkins: they propose a radically different approach, that considers documents simply as symbol sequences, and makes use of specific kernels. The approach is entirely sub symbolic, in the sense that it considers the document just like a unique long sequence, and still it is capable to capture topic information the problem again occur A standard approach to text categorization makes use of the so-called bag of words (BOW) representation, mapping a document to a bag (i.e. a set that counts repeated elements). It losing all the word order information and only retaining the frequency of the terms in the document.

4.7 Effective Pattern Discovery for Text Mining

By Ning Zhong, Yuefeng Li, and Sheng-Tang Wu:they proposed the system An effective pattern discovery technique, is discovered Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patternsSolves Misinterpretation ProbleConsiders 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. 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. But the problem with the system are It is time consuming.Regarding the effectiveness of pattern-based approaches: low frequency and misinterpretation Normally two level patterns present Low Frequency Pattern High Frequncy Pattern.ssMisinterpretation is related to the smeasures used in pattern minin(e.g., “support” and “confidence”). Difficulty is how to use discovered patterns to accurately evaluate the weights of useful features (knowledge) in text documents.

5. EFFECTIVE OPTIMAL SEARCHIG PATTERN MECHANISM USING FOR TEXT MINING

In this propased system ,It will over come all the problem,which will be discovered in last decade,To compared the previous discovered model,and mines concepts of optimal searching using effective discovery of

text mining and relations are more accurately. The traditional text mining generates text database with low-level and the High-level rules are considered as the summary of low level rules. These high level rules are concepts of frequency .Here also used the concepts and their inter relations represented by the ontology are useful for comparing text documents form clusters.Here we will use the concept of offset method .However, the study of text mining based on domain ontology is still in the starting stage, and has not yet formed anorganized study.

Algorithm 1:

```
Input - Positive documents D+
Ouput - Offsets of each term

actualOffset = 0s
EndOfLineOffset = 0

termArray[100]
foreach d in D+ do
//Line ends with \n
foreach Line in d do
Line = RemoveStopWords(Line)
Line = Streaming(Line)
termArray = Line.split(" ")
foreach term in termArray do
UpdateTermInDB(term)
end
end
end

foreach d in D+ do
foreach Line in d do
if Line contains "$$" Then
//This means more than one data set(files) selected for text
mining need to set EndOfLineOffset to 0
EndOfLineOffset = 0
end if
termArray = GetAllTermsFromDB(Line)
foreach term in termArray do
actualOffset = GetTheOffsetFromREGX(term,Line) +
EndOfLineOffset
sUpdateTheOffsetInDB(actualOffset,term,Line)
end
EndOfLineOffset = EndOfLineOffset + Line.length() + 2 //
+2 is for \r\n characters
end
end
```

6.APPLICATION

The main application of the Text Mining are most often used in Text Mining Applications in Natural Language Processing and Multilingual Aspects

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administrationand legal documents.
- Pharmaceutical and research companies and healthcare.

7. CONCLUSIONS

Many data mining techniques have been proposed in the long time. These techniques include association rule of mining, sequential pattern mining, maximum, frequent itemset mining,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 there is very long patterns with high specificity lack in support (i.e., the low-frequency problem). And we argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques organize the ineffective performance of the system. The proposed technique uses four processes, pattern deploying and pattern evolving,suffeling,offset to refine the discovered patterns in text documents. The experimental results show that the proposed model out performs not only other pure data mining-based methods and the conceptbased model.

8.REFERENCES

[1]. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu"Effective Pattern Discoveryfor Text Mining" in,.jan 2012..

[2]. 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.

[3*]* Helena Ahonen Oskari Heinonen Mika Klemettinen A. Inkeri Verkamo "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections.

[4] Hiroki Arimura" Text Data Mining with Optimized Pattern Discovery" Department of Informatics, Kyushu University, Fukuoka 812{8581, Japan PRESTO, Japan Science and Technology Corporationarim@i.kyushu-u.ac.jp

[5]. G. Koteswara Rao1 and Shubhamoy Dey 2**"**DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH "

[6] 1Miloš Radovanovic,2 Mirjana Ivanovic2 "TEXT MINING:APPROACHES AND APPLICATIONS"

[7] Timothy Chklovski and Patrick Pantel. "VERBOCEAN": Mining the Web for Fine-Grained

Semantic Verb Relations by [8]. Text Data Mining with Optimized Pattern DiscoveryHiroki Arimura.

[9] A Survey of Text Mining Techniques and Applications Vishal GuptaLecturer Computer Science & Engineerings University Institute of Engineering & Technology,Panjab University Chandigarh, India

[10]. Masakazu Seno and George Karypis" SLPMiner: An Algorithm for Finding Frequent Sequential PatternsUsing Length-Decreasing Support Constraint" Department of Computer Science and Engineering, Army HPC Research CenterUniversity of Minnesota

[11] Jian Pei Jiawei Han Behzad Mortazavi-Asl Helen Pinto" PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected PatternGrowth" Intelligent Database Systems Research Lab. School of Computing Science, Simon Fraser UniversityBurnaby, B.C., Canada V5A 1S6 E-mail: _ peijian, han, mortazav

[12] 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.

[13] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval(SIGIR '92), pp. 37-50, 1992.

[14] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns withoutCandidate Generation," Proc. ACM SIGMOD Int'l Conf. Managementof Data (SIGMOD '00), pp. 1-12, 2000.

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

[16] X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed SequentialPatterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining(SDM '03), pp. 166-177, 2003.

[17] C K Bhensdadia, Y P Kosta, "An Efficient Algorithm for Mining Frequent Sequential Patterns and Emerging Patterns. with Various Constraints," Machine Learning, vol. 40, pp. 31-60, 2001.

