Efficient Text Mining using Improved IPE Algorithm

[1]Alfiya Sana, [2]Prateek Gupta
[1,2]Dept. of CSE, Shri Ram Institute of Science & Technology, Jabalpur, MP, India

Abstract
Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Over the years, 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. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. For achieving the relevant output which is left due to less frequency, improved IPE has been suggested and proposed to be implemented in this work.

Keywords
Data Mining, Inner Pattern Evaluation, Association Rule Mining, Threshold Based Mining, Pattern Deploying, Pattern Evolving

I. Introduction

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. 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.

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent itemset mining, sequential pattern 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.

The data mining has attracted a great deal of attention in the information technology industry, due to availability of large volume of data which is stored in various formats like files, texts, records, images, sounds, videos, scientific data and many new data formats. There is imminent need for turning such huge data into meaningful information and knowledge. The data collected from various applications require a proper data mining technique to extract the knowledge from large repositories for decision making.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large volume of data. Data mining and knowledge discovery in databases are treated as synonyms, but data mining is actually a step in the process of knowledge discovery. The main functionality of data mining techniques is applying various methods and algorithms in order to discover and extract patterns of stored data. These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base. Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making.

Data mining has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval. Data mining techniques have been introduced to new areas including neural networks, patterns recognition, spatial data analysis, image databases and many application fields such as business, economics, and bioinformatics.

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [1]. Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. A widely accepted formal definition of data mining is given subsequently. According to this definition, data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data [2]. Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is a new term used to identify the datasets that are of large size and have grater complexity [3]. So we cannot store, manage and analyze them with our current methodologies or data mining software tools. Big data is a heterogeneous collection of both structured and unstructured data. Businesses are mainly concerned with managing unstructured data.

The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Mining the information helps organizations to make knowledge driven decisions. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules [4]. It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis [4].

Section I introduced the data mining and its methods applied in industry. Section II discusses the short comings of the data mining techniques. Section III discusses the existing work done in data mining and its short comings. Section IV discusses the

problem statement and Section V elaborates on proposed solution of the problem taken into consideration. Section VI provides the conclusion of the proposed system.

II. Challenges in Mining

Data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

A. Heterogeneity and Incompleteness

The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the properties of the patterns vary greatly.

B. Scale and Complexity

Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, Computer Science & Information Technology (CS & IT) 135 organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analysed.

C. Timeliness

As the size of the data sets to be processed increases, it will take more time to analyse. In some situations results of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. Obviously a full analysis of a user's purchase history is not likely to be feasible in real time. So we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

D. Security and Privacy Challenges

Big data refers to collections of data sets with sizes outside the ability of commonly used software tools such as database management tools or traditional data processing applications to capture, manage, and analyze within an acceptable elapsed time. The extraordinary benefits of big data are lessened by concerns over privacy and data protection. As big data expands the sources of data it can use, the trust worthiness of each data source needs to be verified and techniques should be explored in order to identify maliciously inserted data. Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns. Any security control used for big data must meet the following requirements:

- It must not compromise the basic functionality of the cluster.
- It should scale in the same manner as the cluster.
- It should not compromise essential big data characteristics.
- It should address a security threat to big data environments or data stored within the cluster.

III. Existing System

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, 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. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance [1].

Text mining can be defined as the art of extracting data from large amount of texts. It allows to structure and categorize the text contents which are initially non organized and heterogeneous. Text mining is an important data mining technique which includes the most successful technique to extract the effective patterns. The paper focuses on developing an efficient method for discovering patterns from the document. In text mining field, pattern mining techniques are used to find text patterns, such as frequent item sets, closed frequent item sets, co-occurring terms. This paper presents an innovative and effective pattern discovery technique which includes the process of pattern evolving and pattern deploying, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information [2].

Text mining is a discovery of interesting knowledge in text documents. Exact and accurate knowledge in the text documents needed for the user to find what they require. Many data mining methods are used to mine useful patterns from text documents. However, using and updating these discovered patterns is still an open research issue. Many term based methods are suggested, but a disadvantage with these methods is that they suffer from the problem of synonymy and polysemy. To overcome these disadvantages pattern mining methods are recommended. Pattern mining methods are not proven to be better than term based methods because of low frequency and pattern misinterpretation problem. Here an effective pattern discovery technique is given which applies a pattern co-occurrence matrix to clean close sequential patterns. Process of pattern deploying is applied with the co-occurrence weight and absolute support (PDCS) as deploying approach to overcome pattern misinterpretation problems and pattern evolving to overcome low frequency problem. It also applies a pattern co-occurrence matrix to clean close sequential patterns. This improves performance by using and updating discovered patterns and finding interesting and relevant information [3].

The Internet is a major source of online news content. Current efforts to evaluate online news content, including text, story line and sources is limited by the use of small-scale manual techniques that are time consuming and dependent on human judgments. This article explores the use of machine learning algorithms and mathematical techniques for Internet-scale data mining and semantic discovery of news content that will enable researchers to mine, analyze and visualize large-scale datasets. This research has the potential to inform the integration and application of

data mining to address real-world socio-environmental issues, including water insecurity in the Southwestern United States. This paper establishes a formal definition of framing and proposes an approach for the discovery of distinct patterns that characterize prominent frames. Our experimental evaluation shows that the proposed process is an effective and efficient semi-supervised machine learning method to inform data mining for inferring classification [4].

The burgeoning amount of textual data in distributed sources combined with the obstacles involved in creating and maintaining central repositories motivates the need for effective distributed information extraction and mining techniques. Recently, as the need to mine patterns across distributed databases has grown, Distributed Association Rule Mining (D-ARM) algorithms have been developed. These algorithms, however, assume that the databases are either horizontally or vertically distributed. In the special case of databases populated from information extracted from textual data, existing D-ARM algorithms cannot discover rules based on higher-order associations between items in distributed textual documents that are neither vertically nor horizontally distributed, but rather a hybrid of the two. In this article we present D-HOTM, a framework for Distributed Higher Order Text Mining. Unlike existing algorithms, D-HOTM requires neither full knowledge of the global schema nor that the distribution of data be horizontal or vertical. D-HOTM discovers rules based on higher-order associations between distributed database records containing the extracted entities. In this paper, two approaches to the definition and discovery of higher order itemsets are presented. The implementation of D-HOTM is based on the TMI and tested on a cluster at the National Center for Supercomputing Applications (NCSA). Results on a real-world dataset from the Richmond, VA police department demonstrate the performance and relevance of D-HOTM in law enforcement and homeland defense [5].

With rapid expansion of the numbers and sizes of text repositories and improvements in global connectivity, the quantity of information available online as free-format text is growing exponentially. Many large organizations create and maintain huge volumes of textual information online, and there is a pressing need for support of efficient and effective information retrieval, filtering, and management. Text categorization, or the assignment of textual documents to one or more pre-defined categories based on their content, is an essential component of efficient management and retrieval of documents. Previously, research has focused predominantly on developing or adopting statistical classification or inductive learning methods for automatically discovering text categorization patterns for a pre-defined set of categories. However, as documents accumulate, such categories may not capture a document's characteristics correctly. In this study, we propose a mining-based category evolution (MiCE) technique to adjust document categories based on existing categories and their associated documents. Empirical evaluation results indicate that the proposed technique, MiCE, was more effective than the category discovery approach and was insensitive to the quality of original categories [6].

IV. Problem Statement

The authors have used Inner Pattern Evolution for handling important patterns with low frequency. 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. For IPE a threshold is usually used to classify documents into relevant or irrelevant categories. Using the d-patterns, the threshold can be defined naturally. The IPE uses the set of already evolved patterns showing negativity. This is involving two stages of processing which results in low performance. Also assignment of the threshold values is variant which causes unstable results.

V. Proposed Work

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Over the years, 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. This work presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance. Step by step proposed work is as follows:

Step 1: RCV1 Dataset shall be used to and loaded in the system

Step 2: Pattern Taxonomy using D-Pattern Mining shall be applied to evaluate the patterns and their weights

Step 3: Improved Inner Pattern Evolution shall be applied to select patterns having low frequency but are relevant to the text mining for the users.

Step 4: All the patterns collected in step 2 and step 3 shall be used to classify the documents in RCV1 data set to evaluate TP, TN, FP, FN values

Step 5: TP, TN, FP, FN values of step 4 shall be used to calculate precision, recall, accuracy and F-Measure values.

Step 6: Time taken in every step shall be recorded for measuring the efficiency of the proposed system specially improved IPE.

Step 7: Results and graphs shall be drawn along with the comparison chart of the existing algorithm.

Improved IPE Algorithm (D+, D-)

This will evaluate the average threshold values for the list of negative patterns evaluated after pattern taxonomy. This average will be used to include all the patterns having a greater weight and threshold greater than the average threshold of the negative pattern list.

Improved IPE Algorithm (D+, D-)

```
Begin
        For each Document d in D- do
                x->Evaulate Threshold (d, D-)
                if weight(d)>0 AND x>AvgThread(D-) then
                        include d in pattern list
                end if
        End do
End;
```

VI. Conclusion and Future Work

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset 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 ineffective performance.

References

[1] N. Zhong, Y. Li, S. T. Wu,"Effective Pattern Discovery for Text Mining," In IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, pp. 30-44, 2012.

[2] V. Aswini, S. K. Lavanya,"Pattern discovery for text mining," 2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), Chennai, pp. 412-416, 2014.

[3] R. Gangarde, V. L. Kolhe,"Effective pattern discovery by cleaning patterns with pattern co-occurrence matrix and PDCS deploying approach," First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, pp. 119-124, 2014.

[4] L. H. Cheeks, T. L. Stepien, D. M. Wald,"Discovering News Frames: Exploring Text, Content, and Concepts in Online News Sources to Address Water Insecurity in the Southwest Region," 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA, USA, pp. 454-462, 2016.

[5] S. Li et al.,"Mining Higher-Order Association Rules from Distributed Named Entity Databases," IEEE Intelligence and Security Informatics, New Brunswick, NJ, USA, pp. 236-243, 2007.

[6] Chih-Ping Wei, Yuan-Xin Dong,"A mining-based category evolution approach to managing online document categories," Proceedings of the 34th Annual Hawaii International Conference on System Sciences, Maui, HI, USA, pp. 10, 2001.