International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing(ICIAC- 12th & 13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Pattern Taxonomy Model used in Text Mining with respect to different Application Domains

Ms. Priya P. Malkhede[#1], Ms. Hemlata Dakhore[*2]
[#1]M. Tech. Student, CSE Dept., GHRIETW/RTMNU, India
[1]priyapmalkhede@gmail.com
[*2]Assistant Professor, CSE Dept., GHRIETW/RTMNU, India
[2]hemlata.dakhore@raisoni.net

Abstract
In the last decade, many data processing techniques are planned for mining helpful patterns in text documents. However, how effectively use associate degreed update discovered patterns remains an open analysis issue, particularly within the domain of text mining. Existing system is used term-based approach for extracting the text. Pattern evolution technique is used to improve the performance of term-based approach. The term-based approach is suffered from the problems of polysemy and synonymy. Following methods gives a way to boost the effectiveness of victimization and change discovered patterns for locating relevant and attention-grabbing data. Substantial experiments on information assortment and topics demonstrate that the planned answer achieves encouraging performance. Researchers are still going in efficient updating of discovered pattern.

Keywords— Text Mining, Information Retrieval, Pattern Taxonomy Model, Pattern Deploying, Pattern evolving, sequential pattern mining.

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 a forthcoming need for turning such data into useful information and knowledge [1]. Valuable information is always needed in all sort of information extraction. Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. It is therefore crucial that a good text mining model should retrieve the information that meets users' needs within a relatively efficient time frame. Traditional Information Retrieval (IR) has the same goal of automatically retrieving relevant documents as many as possible while filtering out non-relevant ones at the same time. However, IR-based systems cannot meet users' needs. The unnecessary information can be removed with this proposed work. In today's world the whole system is totally depended on computer system. The data stored in system can be of various types like text, video, audio, rich document and so on. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [4].

Traditionally, texts have been analyzed using various information retrieval related methods, such as full-text study, and natural language processing. Recently, we have seen the high-spirited appearance of very large heterogeneous full-text document collections, available for any end user. The variety of user's wishes is broad. The user may need an overall view of the document collection: what topics are covered, what kind of documents exist, are the documents somehow related, and so on. On the other hand, the user may want to find a specific piece of information content. At the other extreme, some users may be interested in the language itself, e.g., in word usages or linguistic structures. The aim of this research is to make a significant contribution in dealing with the information mismatch and overload problems.

Most research work in data mining field focuses on developing efficient mining algorithm for discovering a variety of patterns from a larger data collection [5]. However searching for useful and interesting patterns is still open research issue. In the field of text mining, data mining techniques can be used to find various text patterns, such as association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining.

II. METHODOLOGIES USED IN TEXT MINING

A. Knowledge Discovery

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have

created an enormous need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions. Knowledge discovery consist of following steps: data selection, data preprocessing, data transformation, pattern discovery and pattern evaluation.

1) Data Selection: This process includes generating a target dataset and selecting a data set or a subset of large data sources where discovery is to be performed. The input for this process is a database and output is aim data.

2) Pre-Processing: This process involves data cleaning and noise removing. It also includes collecting required information from selected data fields, providing appropriate strategies for dealing with missing data and accounting for redundant data.

3) *Transformation:* The preprocessed data needs to be transformed into a predefined format; depending on the data mining task. This process needs to select an adequate type of features to represent data. In addition, feature selection can be used at this stage for dimensionality reduction. At the end of this process, a set of features is recognized as a dataset. All title and author details must be in single-column format and must be centred.

4) *Data Mining:* Data mining is a specific activity that is conducted over the transformed data in order to discover patterns. Depend on user requirements, the discovered patterns can be pairs of features from the given dataset, a set of ordered features occurring together, or a maximum set of features.

5) *Evaluation:* The discovered patterns are evaluated if they are valid, novel and potentially useful for the users to meet their information needs. Only those evaluated to be interesting in some manner are viewed as useful knowledge. This process should decide whether a pattern is interesting enough to form knowledge in the current context.

B. Association Rules

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. It helps to find out frequently co-occurring elements by analyzing the data. The two common measures of rule interestingness or usefulness are support and confidence. This rule is used to discover the associatively between objects. The problems of mining association rules from large databases can be decomposed into two sub problems. (1) Find item sets whose support is greater than the user specified minimal support. (2) Use the frequent item sets to generate the desired rules.

C. Sequential Pattern Mining

Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. Sequential pattern mining discovers subsequences that are common to more than minsup sequences in a sequence database, where minsup is set by the user. A sequence is an ordered list of transactions.

D. Frequent Itemsets

Frequently occurring subsets in a sequence of a set is frequent itemset. Many algorithms were proposed for extracting the frequent itemset. Apriori is the widest known algorithm. Later many apriori modifications were proposed, DHP (Direct Hashing Pruning), DIC (Direct Itemset Counting), sampling and partition [6].

III. TEXT CLASSIFICATION

Text classification is the process of classifying documents into predefined categories based on their content. It is the automated assignment of natural language texts to predefined categories. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data. Existing supervised learning algorithms to automatically classify text need sufficient documents to learn accurately.

A. Decision Tree

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making [6]. The goal is to create a model that predicts the value of a target variable based on several input variables.

B. K-nearest Neighbour

K-nearest-neighbor (KNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.

C. Naïve-Bayesian Approach

A Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [3]. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

IV. TEXT REPRESENTATION

A. Keyword-Based Representation

The bag-of-words scheme is a typical keyword based representation in the area of information retrieval. It has been widely used in text classification due to its simplicity, but this method causes the ambiguity. In IR-related tasks, if a query contains an ambiguous word, the retrieved documents may have this word but not its intended meaning.

B. Phrase-Based Representation

Using single words in keyword-based poses the semantic ambiguity problem. In general, phrases carry more specific content than single words. Another reason for using phrase-based representation is that the simple key-word based representation of content is usually inadequate because single words are rarely specific enough for document discrimination. To identify group of words that create meaningful phrases is a better method, especiallyfor phrases indicating important concepts in the text. There are five categories of phrase or terms extraction: (1) Co-occurring terms, (2) Episodes, (3) Noun phrase, (4) Key-Phrase, (5) ngram. The drawback of phrase based representation is given in [8].

C. Pattern Based Approach

There are two main considerations regarding efficiency of pattern based approach, which is low frequency and misinterpretation. If the minimum support is decreased a lot of noisy patterns can be found. It means that whatever decided by the end user as a desired output that will be not given by your system.

V. PATTERN TAXONOMY MODEL

Two Main stages are considered in Pattern Taxonomy Model. The first stage is how to extract useful phrases from text document. The second is the how to use these discovered patterns to improve the effectiveness of a knowledge discovery system. As a first step in this work the given text documents are separated into different paragraphs. So treat each paragraph as an individual transaction, which consist of a set of words (terms) at the subsequent phase apply the data mining method to find frequent patterns from these transaction and generate pattern taxonomies. During this process meaningful and redundant patterns are eliminated by applying proposed pruning scheme. Positive documents are the documents that are frequently occurring in the database. Here considering only the positive documents. For the classification into positive documents we are using the Naïve Bayesian Algorithm. This model follows two steps in first step it describes in what way extract the patterns from the text documents in second step it describes How to update the discovered patterns effectively for performing the knowledge discovery from the text documents in PTM . In this technique we split the document into a paragraphs and each paragraph is to be taken as a one document .Let us assume a given document is considered as d and it yields PS (d).

Stop word removal: In computing, stop words are words which are filtered out prior to, or after, processing of natural language data (text).there is not one definite list of stop words which all tools use and such a filter is not always used some tools specifically avoid removing them to support phrase .Any group of words can be chosen as the stop words for a given purpose [1]. For some search machines, there are some of the most common, short function words, such as the, is, at, which and on.

Stemming Process: Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms. In this preprocess the text documents have to be processed using the Porter stemmer. It removes the Suffix's of the words these words are useful in the text mining for clustering the text documents in the text mining process we collects the documents and each documents are composed into the set of terms or words. The words having stem have a same meaning .in stem process the suffixes of the words, singular and plural words are considered into a one single word for meaning full text clustering

process [1]. After the preprocessing the text documents will giveas an input to PTM.

A. Pattern Pruning

For all methods used for finding all frequent sequential patterns from a data set, the problem encountered is large amounts of patterns are generated. Most of which are considered as no meaningful patterns and need to be eliminated. A proper pruning scheme can be used for addressing this problem by removing redundant patterns, and decreasing the effects from noise patterns. For that purpose the proposed work use closed patterns as meaningful patterns since most of subsequence patterns of closed patterns have the same frequency, which means they always occur together in a document.

B. Using Discovered Patterns

The SPMining uses the sequential data mining technique with a pruning scheme to find meaningful patterns from text documents but SPMining not overcome the issue of how to use these discovered patterns. There are various ways to utilize discovered patterns by using a weighting function.

C. Evaluation of Discovered Patterns

In This Phase two methods will use they are Deployed Pattern Evaluation (DPE) and In Pattern Evaluation (IPE). This model will test how to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. This technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem so; this technique is called inner Pattern evolution, because it only changes a pattern's term supports within the pattern.

VI. PATTERN DEPLOYING METHOD

The properties of pattern (e.g. support and confidence) used by data mining methods in the phase of pattern discovery are not suitable to be adopted in the phase of using discovered patterns. Pattern deploying method consist of more terms is considered to be more specific but its frequency is very low. This method is proposed for solving the problem caused by inappropriate evaluation of patterns, discovered using data mining methods such as SPM (Sequential Pattern Mining) and NSPM (Non-sequential Pattern Mining) utilize discovered patterns directly without any modification and thus encounter the problem of lacking frequency on specific patterns.

VII. SIGNIFICANCE

This study will contribute to current efforts in establishing better systems which gives us the way how effectively deal with the large amount of discovered patterns as well as solve the problem of redundancy in database for polysemy word and synonymy word.This approach is useful to reduce the time required to discover the pattern and improves the accuracy. For this challenging issue[1] used, closed sequential patterns for text mining which gives that the concept of closed patterns in text mining useful for interpretation of discovered patterns in text documents and had the potential for improving the performance of text mining.

VIII. CONCLUSIONS

This paper gives many data mining concepts and techniques which are helpful in solving many problems. These techniques include association rule mining, frequent item set 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 therefore, this proposed approach motivated the field of research, gave a more formal definition of the terms used herein and presented a brief overview of currently available text mining methods, their properties and their application to specific problems. A PTM model with new pattern discovery model for text mining mainly focuses on the implement of temporal text pattern. The prime aim of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. But those methods working based on the term support that methods are ignoring the relationships between the terms. 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.

REFERENCES

[1] Ning Zhong, Yeufeng Li and Sheng-Tang Wu, "Effective Pattern Discovery fot Text Mining," IEEE Trans. Knowledge and Data Engg., vol. 24, No. 1, Jan 2012.

[2] Chih-Ping Wei and Yu-Hsiu Chang,"Discovering Event Evolution Patterns From Document Sequences,"Published in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 37, NO. 2 MARCH 2007.

[3] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification,"Published in *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 11*, NOVEMBER 2006.

[4] Kavitha Murugeshan, Neeraj RK, "Discovering Patterns to Produce Effective Output through

Text Mining Using Naïve Bayesian Algorithm,"*Published in International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-6,* May 2013.

[5] Mrs. K. Yasodha, Research Scholar, Mrs.K. Mythili, Professor, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining" *Published in International Journal of Science and Applied Information Technology, Volume 1, No.3,* July – August2012.

[6] Anisha Radhakrishnan, Mathew Kurian,"Efficient Updating of Discovered Patterns for Text Mining: A Survey," *Published in International Journal of Computer Applications* volume no 58 – No 1, November 2012.

[7] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member IEEE, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words," *Published in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA, VOL. 23, NO. 7,* JULY 2011.

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

[9] 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.