International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Improved Prototype Text Mining for Data Knowledge Discovery

Varikuti. Srividya[1], Pathuri SivaKumar[2]

[1]Computer Science and Engineering, Rise Group of Institutions, Ongole, India

Abstract: ***Digital data in the form of text documents is rapidly growing. Analyzing such data manually is a tedious task. Data mining techniques have been around to analyze such data and bring about interesting patterns. Many existing methods are based on term-based approaches that can't deal with synonymy and polysemy. Moreover they lack the ability in using and updating the discovered patterns. Zhong et al. proposed an effective pattern discovery technique. It discovers patterns and then computes specificities of patterns for evaluating term weights as per their distribution in the discovered patterns. It also takes care of updating patterns that exhibit ambiguity which is a feature known as pattern evolution. In this paper we implemented that technique and also built a prototype application to test the efficiency of the technique. The empirical results revealed that the solution is very useful in text mining domain.***

Keywords: Text mining, Pattern mining, Sequential pattern mining, closed frequent mining, Pattern taxonomy, Information retrieval.

1. Introduction

Knowledge discovery has become an indispensable phenomenon in recent years due to the rapid increase in digital data. They have attracted lot of attention in academic and scientific circles. Many applications in the real world need such mining of data in order to discover trends or patterns. These trends or patterns lead to business intelligence (BI). Such BI helps in taking well informed decisions. Many data mining techniques came into existence in the past ten years. They include closed pattern mining, maximum pattern mining, sequential pattern mining, item set mining, and association rule mining. These techniques are developed for data mining algorithms. They are capable of producing huge number of patterns. However, how to use those patterns and how to update them in future is the area that needs some more research. Especially in the field of text mining, patterns are discord from text documents. It is a challenging job to use those patterns and also update them. Earlier term based methods are provided by Information Retrieval (IR) techniques. The term based methods are classified into rough set models, SVM based models and probability models. All the term based methods suffer from problems such as synonymy and polysemy. When award has many meanings it is known as polysemy. When multiple words have similar meaning, it is called synonymy. Thus the discovered patterns with term based techniques have semantic meaning and answering the exact user query is difficult.

For this reason for many years people started believing that phrase-based techniques are better than that of term – based. However, the experiments in the field of data mining have not been proved. The possible reasons include the phrases have less properties pertaining to statistics when compared with terms; frequency of occurrence is low; noisy and redundant phrases are more.

Though there are some drawbacks, the sequential patterns became promising alternatives to phrase. The reason for this is that sequential patterns avail required statistics like terms. Pattern Taxonomy Models (PTMs) came into existence to overcome the drawbacks of phrase-based mining approaches. Pattern based approaches became alternatives but much improvements are not made to make them more effective for text mining. With regard to effectiveness there are two issues. They are misinterpretation and low frequency. When patterns are less frequent, they can't be used for decision making. When the terms or patterns are misinterpreted, the result will not be reliable. Low frequency can't have required support. If the support is decreased, the results may not be useful for business decisions.

Over the last many years Information Retrieval (IR) is also used to have many techniques that used features of text documents. They are used to retrieve content from huge amount of documents based on the terms and their weights. The terms may have different weights based on the context as well. There might be semantic meanings that are to be considered in IR. Therefore it is not sufficient to only consider weights of terms for document analysis or evaluation. In this paper we implement a novel pattern discovery technique proposed by Zhong et al. It first computes specificities of the discovered patterns and then evaluates the weights of terms based on the distribution. Thus it is capable of avoiding misinterpretation problem. Negative training examples influence is also considered by this in order to avoid low frequency problem. Moreover the ambiguous patterns are updated. This phenomenon is known as pattern evaluation. Thus the proposed approach improves accuracy of the discovered patterns.

2. Ease of Use

In basic technique, the information retrieval provides a term-based method which consists of terms or words to find an approach to effectively use and update the discovered patterns from the text documents. The termbased method suffers from the problem 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 finding what users want. To overcome the issue of term-based method, a pattern-based approach was developed in which patterns are a set of terms and used to extract the patterns from large amount of text documents. There are two

Licensed Under Creative Commons Attribution CC BY

fundamental issues in pattern-based approach: low frequency and misinterpretation. Low frequency occurs when the minimum support value is decreased and small patterns are discovered. Misinterpretation is the measure used in pattern mining that is not suitable to discover the patterns. Ning Zhong, Yuefeng Li, And Sheng-Tang Wu, [1] provides an innovative and effective pattern discovery technique which includes the process of pattern taxonomy, pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. It uses the term based approach which suffers from the problem of polysemy and synonymy and phrase based approach which takes only the semantic information and they are less ambiguous and discriminative then individual terms. The semantic meaning is not useful for answering what the user's want. Dipti S.Charjan and Mukesh A. Pund, [2] provides an efficient mining algorithm for discovering patterns from large data collection and also searches for the interesting patterns. It mainly focuses on the pattern repetitions and investigates the better means of long patterns. It also supports the filtering of patterns for user's usage purpose.S.Habeeba and S.Nawaz [3] implements the temporal text mining approach and extracts the sequence of events from the news and other documents based done the published time of documents. The technique is enhanced to web search through processed text mining. Inje Bhushan.V and Ujwalapatil [4] present a survey on the different pattern mining techniques to extract the patterns as per the user's need. K.Rupika and lSharmiliya Fransuva and M.Suganthy, [5] provides the concept of text categorization techniques to update and use the discovered patterns form the text mining. It includes the pattern taxonomy model, pattern deploying and IP Evolving to identify the patterns from the document. Sabarina.K, [6] provides an effective pattern discovery to boost the effectiveness and also to change the discovery of patterns for identifying the relevant data. It is used to overcome the problems of misinterpretation and low frequency and outperforms the pure data and also term based models. D.Shanthi and V.Umadevi, [7] provides a discussion on pattern based approach with patten extraction techniques and compares the result with models on extact sequence quality and execution time. It uses the sequential patterns to extract the patterns and improves the space and time of the used algorithms. Rohini Y.Thombare and Shirish.S.Sane [8] suggest a probabilistic method to estimate the accurate term weights and the support values to extract the patterns. Sujit V.Chaudhari and Shrikant Lade, [9] utilize the patternbased approach with the set of keywords and classify the documents related to research articles and news articles. It provides an effective pattern form relevant or positive documents and a set of keywords that contains the relevant fields. Anisha Radhakrishnan and Mathew Kurian, [lO] provides an approach to use and update the patterns to find relevant information from the text document by using two methods pattern evolving and deploying. The updating of discovered pattern effectively was difficult with these techniques because the long patterns with high specificity lacks in support. K. Mythili, and K. Yasodha, [11] introduces the implement of temporal text mining approach. It focuses on knowledge and loss less decomposition algorithm to analyze the relationship between the assigned time period of text document and the information of the document fortemporal analysis. A.Anil Kumar and S.Chandrasekhar,[12] presents an adequate pre-processing and dimensionality reduction techniques which eliminates the noisy data, identifies the root word and reduces the size of text data. It uses a singular value decomposition technique to reduce the dimension of data. To retrieve the relevant information, clustering process is used. Sayantani Ghosh, Sudipta Roy, and Prof. Samir K. Bandyopadhyay, [13] provides different text mining algorithms that involves the process of structuring the input text, discover patterns, and evaluates the output. It includes the process of different techniques such as information retrieval, natural language processing, and information extraction. Kanak Saxena and D.S.Rajpoot, [14] provides an idea to extract patterns from collection of data and also describes the classification techniques. Gerard Salton and Christopher Buckley, [15] provides an analysis of assigning the automatic term weights and compares the term indexing models with the content analysis. It mainly focuses on the effectiveness of data retrieval with two main factors such as the items that are relevant based on the user's need must be retrieved and other items are rejected. Vidhya.K.A and G.Aghila, [16] provides a survey on the text classification which uses the naive bayes approach toclassify the large datasets and the working mode of the algorithm. It also surveys the feature selection method and different text document classification metrics.

3. Proposed Scheme

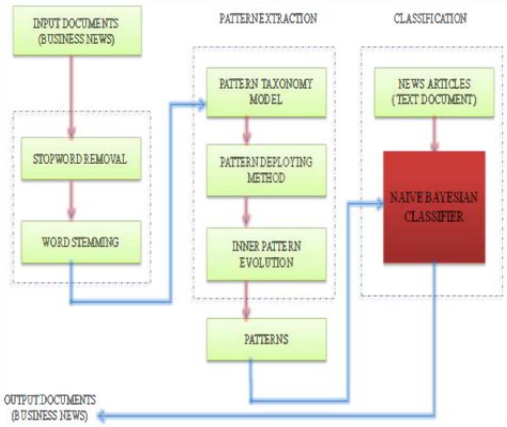

Fig 1: System Architecture

In proposed scheme, the pattern-based approach is used in which, the discovered patterns are more specific than the whole documents. The patterns are a set of terms extracted from the documents. The dataset used for the present work is business news from "THE HINDU" e-newspaper and it is stored in .txt file. The news document is given for pre-processing techniques which holds stop word removal and stemming to reduce the useless words that have no meaning. The word-list from pre-process technique is given to pattern extraction process which includes pattern taxonomy model (PTM) to extract the patterns. It includes pattern deploying method and Innerpattern evolution to evaluate the terms weights and reduce the noisy patterns from the text documents. Each pattern includes the set of terms that are extracted from the pattern taxonomy model in which the frequent-closed patterns and sequential patterns are performed. Fig. 1 describes the overall design of the work.

The proposed scheme includes five modules:

Licensed Under Creative Commons Attribution CC BY

3.1 Text Pre-processing

A database consists of large amount of data which are collected from different sources of data. If the data is inconsistent, then the mining process can lead to confusion which results in inaccurate data. In order to extract the accurate and consistent data, data pre processing is applied. The main objective of pre processing is to obtain the key features or key terms from news text documents and to enhance the relevancy between the word and document. It is important to select the keywords that carry the meaning, and discard the words that do not contribute to distinguish between the documents. The text pre-processing includes two techniques: Stop-Word Removal and Word Stemming

3.1.1. Stop-Word Removal

The most common words in any text document that does not provide any meaning of the documents are prepositions, articles and pronouns. These words are treated as stop-words. Every document deals with these words which are not necessary for text mining applications and hence these words are eliminated. Example of such words are 'the',' in',' a',' an' etc.

3.1.2 Word-Stemming

Stemming or lemmatization is a technique that reduces the words into their root or stem. The hypothesis of stemming is that the word with the same stem or root describes the same or relatively close concepts in the text document. Example, agree, agreed, agreeing, agreement belong to root word 'agree'. In present work, Porter stemming algorithm is used to find the root words in the document.

3.2 Pattern Taxonomy Model

In PTM method, it includes two main stages, first is how to extract the words from the text documents and second is how to use the discovered patterns to improve the effectiveness of a knowledge discovery system. In present work we assume that all the documents are split into paragraphs and each paragraph is treated as an individual document which consists of set of terms. Let D be a training set of documents. Let T= $\{t_1, t_2, t_3 \ldots \ldots t_6\}$ be a set of terms which are extracted from the documents. Algorithm 1 describes the process of extracting the frequent patterns.

3.2.1. Frequent and Closed Patterns

Frequent-Pattern mining finds a set of patterns that occur frequently in a data set, where a pattern can be a set of terms, a subsequence, or a substructure. A pattern is considered frequent if its count satisfies a minimum support. Given a termset X in document d, d, X is used to denote the covering set of X for d, which includes all paragraphs dp ε PS(d) such that X ⊆dp, i.e.

$$[X] = \{dp | dp \in PS(d), X \subseteq dp \quad (1)$$

To extract the frequent patterns the minimum support is assigned and the support value is calculated based on absolute support (sup_a) and relative support (sup_r).The absolute support is the number of occurrence of terms (X) in PS(d), i.e. sup_a (X) = |X| and the relative support is the fraction of paragraphs that contain the pattern, i.e.

$$sup_r(X) = \frac{|X|}{|PS(d)|} \quad (2)$$

A term set X is called frequent pattern if its sup_r (or) sup_a ≥minimum support.

Algorithm 1 Pattern Extraction:

Input: Set of Word List
Output: Set of frequent patterns.
1: **begin**
2: Documents (d) →split into paragraph (PS)
3: **for each document**
4: Calculate the absolute support (sup_a) and relative support (sup_r)
5:assign the min_support value
6: **if** (sup_a **or** sup_r ≥ **min_sup**)
7: compute the frequent patterns
8:eliminate the short patterns which are relevant to long patterns
9: **end if**
10: **end for**

Table 1 lists the set of paragraphs for a document d, in which PS (d) = {dpb dp2, dp3 ... dp6} and other irrelevant terms are removed. Based on the min_sup value, the frequent patterns and its covering sets in table 2 are obtained. All frequent patterns are not useful since the shorter pattern is a part of longer patterns. A pattern is close if no one of its superset has the same support value.

Table 1: Set of Paragraphs

Paragraph	Terms
dp_1	t_1 t_2
dp_2	t_3 t_4 t_6
dp_3	t_3 t_4 t_5 t_6
dp_4	t_3 t_4 t_5 t_6
dp_5	t_1 t_2 t_6 t_7
dp_6	t_1 t_2 t_6 t_7

Licensed Under Creative Commons Attribution CC BY

Table 2: Frequent Patterns and Covering Sets

Frequent Pattern	Covering Set
$\{t_3, t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2\ dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_1, t_2\}$	$\{dp_1\ dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_6\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_6\}$

3.2.2. Closed Sequential Patterns

Closed sequential pattern is a frequent sequential pattern such that it is not included in any other sequential pattern having exactly the same support.

A sequential pattern S = $<t_{1.....t_r}>$ ($t_i\ \varepsilon\ T$ T) is an ordered list of terms. A sequence S, = $<X_1, X_2, \ldots X_i;>$ is a subsequence of another sequence S2 = $<Y_1, Y_2, \ldots Y_j>$ is called S, is subset of S2, if and only if it belongs to $j_1, j_2 \ldots j_y$ such that $1 \leq j_1 \leq j_2 \leq j_y \leq j$ and $X_i = Y_{j1}\ X_i = Y_{jy}$. Given S_1 is sub-set of S_2 then S, is a sub-pattern of S_2, and S_2 is a super pattern of S, and hence it is said to be sequential patterns. A sequential pattern X is called frequent pattern if its relative support (or absolute support) ≥2: min_ sup, a minimum support. A frequent sequential pattern X is called closed if not any super pattern x I of X such that sup_a (X1) = sup_a (X).

4. Conclusion and Future Work

Many data mining techniques have been proposed in the last decade for mining useful patterns relevant to what user's want. It includes association rule mining, sequential pattern mining, maximum and closed pattern mining. However, using the discovered pattern in the field of text mining is difficult and ineffective, because useful long patterns with high specificity lack in support. In this work, we have focussed on the discovering the patterns from the large amount of data and search for interesting patterns that user want. In proposed technique, the pattern taxonomy model is used to extract the pattern from the documents. The pattern includes the terms of business news articles. The method has been successfully implemented and patterns are extracted correctly based on the documents.

In our future work, we will work with pattern deploying and inner pattern evolution to reduce the noisy patterns and also to evaluate the term weights. We also propose a classification algorithm, called naive bayesian classifier in which the extracted patterns are considered to be training set and news articles as testing set are classified and the business news patterns are extracted from the documents

References

[1] Ning Zhong, Yuefeng Li and Sheng-tang Wu "Effective Pattern Discovery For Text Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. I, January 2012.
[2] Dipti S.Charjan and Mukesh A.Pund "Pattern Discovery for Text Mining Using Pattern Taxonomy", International Journal of Engineering Trends and Technology (UETT), Volume 4 Issue 10- October 2013.
[3] S.Habeeba and S.Nawaz "Discovery of Noisy Patterns in Text Mining," International Journal Research in Computer and Scientific Technology, August 2013.
[4] Inje Bhushan.V and Ujwalapatil,"A Comparative Study on Different Types of Effective in Text Mining: A Survey", International Journal of Computer Engineering and Technology (IJCET), March-April 2013.
[5] M. Suganthy, K.Rupika and J.Sharmiliya Fransuva "Text Mining for Pattern Identification," International Journal of Futuristic Science En gineering and Technology, Volume I Issue 3 March2013.
[6] Sabarina.K, "Application of Pattern Mining Algorithm for Text mining", UREAT International Journal of Research in Engineering & Advanced Technology, Volume I, Issue 1, March, 2013.
[7] D.Shanthi and V.Umadevi, "Mining the Text Documents Using Phrase Based Tokenize Approach", International Journal of Engineering Science Invention, January 2013.
[8] Ning Zhong, Yuefeng Li and Sheng tang Wu "Effective Pattern Discovery For Text Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. I, January 2012.
[9] Sujit V. Chaudhari and Shrikant Lade "Classification of News and Research Articles Using Text Pattern Mining," TOSR Journal of Computer Engineering (IOSR-JCE), Volume 14 Issue 5, 2013.
[10] Mathew Kurian & Anisha Radhakrishnan "Efficient Updating Of Discovered Patterns For Text Mining: A Survey", International Journal Conference, November 2012.
[11] Mythili, and K. Yasodha, Research Scholar, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining", International Journal of Science and Applied Information Technology, August 2012 ..
[12] JA.Anil Kumar and S.Chandrasekhar 'Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering," International Journal of Engineering Research and Technology (UERT), Volume I Issue 5, July 2012.
[13] JSayantani Ghosh, Sudipta Roy, and Samir K. Bandyopadhyay, "A Tutorial Review on Text Mining Algorithms" International Journal of Advanced Research in Computer and Communication Engineering Vol. I, Issue 4, June 2012.
[14] Kanak Saxena and D.S.Rajpoot "A Way to Understand Various Patterns of Data Mining Techniques for Selected Domains," International Journal of Computer Science and Information Security, Volume 6 2009.
[15] Gerard Salton and Christopher Buckley, "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, 1988.

Licensed Under Creative Commons Attribution CC BY