Indian Journal of Research in Applied Sciences Engineering (IJRASE)
Vol.2.No.1 2014pp 57-63
available at: www.goniv.com
Paper Received :05-03-2014
Paper Published:28-03-2014
Paper Reviewed by: 1. John Arhter 2. Hendry Goyal
Editor : Prof.Dr. G.Ramesh

A Novel and Effective Pattern Discovery Technique for Text Mining

T. Balasubramanian
Assistant Professor
Sri Vidhya Mandir Arts & Science College
Katteri, Uthangakari -635 307.

Abstract

Many data processing techniques are proposed for mining helpful patterns in text documents or contents of files. However, how to efficiently use and keep posted those discovered patterns remains an open analysis issue, particularly within the domain of text mining. The common existing text mining ways adopted term-based approaches; all of them suffer from the issues of vagueness and synonymous. Here in this work presents innovative, novel and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of victimization and change discovered patterns for locating relevant and conspicuous data. This system is implemented in MATLAB Text mining tool.

Keywords: Text Mining, Pattern Mining, Pattern Taxonomy, Pattern Evolution.

1. Introduction

Text mining[1] is similar to data mining concepts, except that data mining tools [2] are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, documents, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies and very big organization. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is understandable: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the capability to distinguish and apply linguistic patterns to text and humans can easily overcome difficulties that computers cannot easily handle and understand such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to understand unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Figure 1 on next page, depicts a generic process model [3] for a text mining application.

Text mining is the technique that helps to users finds useful information from a large amount of digital text data [3]. It is therefore critical that a good text mining model should regain the information that users require with relevant efficiency. Traditional Information Retrieval

(IR) has the same objective of automatically retrieving as many relevant documents and files as possible at the same time as filtering out irrelevant documents at the same time. However, IR-based systems do not sufficiently provide users with what they actually need. Many text mining methods have been established in order to achieve the goal of retrieving for information for users. Here focus on the development of a knowledge discovery model to efficiently use and update the discovered patterns and apply it to the field of text mining. The process of knowledge discovery may consist as following:

• Data Selection
• Data Processing
• Data Transaction
• Pattern Discovery
• Pattern Evaluation.

Text mining is also called as knowledge discovery(KD) in databases because, it is frequently find in literature text mining as a process with series of partial steps among other things also information extraction as well as the use of data mining. It is analyze data in knowledge discovery (KD) in databases is aims of finding hidden patterns as well as connections in those data. While the ability to search for keywords or phrases in a collection are now widespread such search only slightly supports discovery because the user has to decide on the words to look for. On the other hand, text mining results can suggest "interesting" patterns to look at, and the user can then accept or reject these patterns as interesting. In this
work here it is deliberated pattern taxonomy model which extracting evocative frequent patterns by pruning the worthless ones. Patterns are sorted based on their reparations.

II. Text Mining Process

There are almost five major technique categories in the text mining process: document retrieval, data extraction, data preprocessing (cleansing), data analysis, and data visualization.

A. Document Retrieval

Information (or document) retrieval is a discipline concerned with the organizing, storage, searching, and retrieval of bibliographic information. It is a powerful framework for analyzing and structuring documents. The text mining model procedures can be divided into three stages: document indexing, term weighting, and computation of similarity coefficients.

B. Data Extraction

Data extraction is the activity of automatically pulling out appropriate information from large volumes of text information. Extraction can take two forms; one is to identify the specific field of entity extracted such as name, date, or address, and the other one is to identify the parts of speech from text corpus using natural language processing (NLP) technology. It employs a combination of semantic and syntactic analyses. It processes text inputs as follows.

• Differentiates and separates each sentence.
• Applies lexicon analysis to categorize nouns, verbs, etc., based on the underlying dictionary.
• Refines word attribution based on syntactic inferences.

This then tags each word with the part(s) of speech it is likely to be.

C. Data Preprocessing

Data preprocessing, or data cleansing, is the algorithm that detects and removes errors or inconsistencies from data and fuses similar data in order to improve the quality of succeeding

analyses. This cleaned data will then be fed to the analysis process. Several methods could be used to clean the data. Three methods that are used in here cleanup are stemming algorithm, elemental fuzzy logic to consolidate like terms, and thesauri. Word stemming or truncation can be used to achieve a quick approximation to the word or text root. A word for which one wants to find an exact or near match may by written as a stem or root word, and the repossession system asked to find words that match the root. One approach used to determine the root of a word is to determine the semantic root such that "box" and "boxes" are equal. Fuzzy matching techniques can be used to identify, associate, and reduce data appropriately. For example, this will handle misspellings, alternative hyphenation and capitalization. A vocabulary is defined as a grouping of terms, into a certain concepts. This can be used for specialized data reduction.

D. Data Analysis

As stated before, each document can be represented as a vector in a high dimensional space. Hence, dimensionality reduction techniques are required to represent n-dimensional document data by a small number of noteworthy dimensions. There are several techniques that have been used for dimensionality reduction, including Factor Analysis (FA)/Principal Component Analysis (PCA) and Cluster Analysis.

E. Data Visualization

A general goal of analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) among the investigated objects. This is accomplished by solving a minimization problem such that the distances among points in the conceptual low-dimensional space match the given (dis)similarities as closely as possible. In factor analysis, the similarities among objects (e.g., terms) are expressed in the correlation matrix. With MDS, one may analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices. However, a major weakness of MDS is that there are no quick and fast rules to interpret the nature of the resulting dimensions.

Problem Statement

Most research works in the data mining community have focused on developing well-organized mining algorithms for discovering a variety of patterns from a larger data collections. However, searching for useful and interesting patterns is still an open problem. In the field of text mining, data mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets, co-occurring terms and multiple grams, for building up a illustration with these new types of structures [19]. On the other hand, the first problem is how to effectively deal with the huge amount of patterns produced by using the data mining methods.

Using phrases for the text illustration immobile has doubts in increasing recital over domains of text categorization tasks , meaning that there exists no particular illustration method with dominating advantage over others [8, 12]. Instead of the keyword-based approach which is typically used by text mining-related tasks in the past, the pattern-based model (single term or multiple terms) is employed to perform the same concept of task. There are two phases that here it is need to consider when we use pattern-based models in text mining: one is how to discover useful patterns from digital text documents, and the other is how to exploit these mined patterns to improve the system's performance.

II. Related Work

A. Pattern Taxonomy

Patterns can be structured into a taxonomy by using there is a (or subset) relation. For the example of Table 1, where we have illustrated a set of paragraphs of a document, and the discovered 10 frequent patterns in Table 2 if assuming min sup ¼ 50%. There are, however, only three closed patterns in this example. Explains the Pattern Taxonomy in Details

1. An outcome of pattern discovery technique is discovered.
2. The proposed approach can advance the good accuracy of calculating term weights because revealed patterns are more ambiguous than whole documents.
3. The process of updating ambiguous patterns can be referred as pattern evolution.
4. Solves Misunderstanding Problem
5. Considers the inspiration of patterns from the negative training examples to find vague (noisy) patterns and tries to reduce their influence for the low-frequency problem.
6. Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns.
7. The incoming documents then can be sorted based on these weights.
8. In training phase the d-patterns in positive documents (Dþ) based on a minimum support are found, and evaluates term supports by deploying dpatterns to terms
9. In Testing Phase (TP)to revise term supports using noise negative documents in D based on an experimental coefficient

Advantages of proposed system:

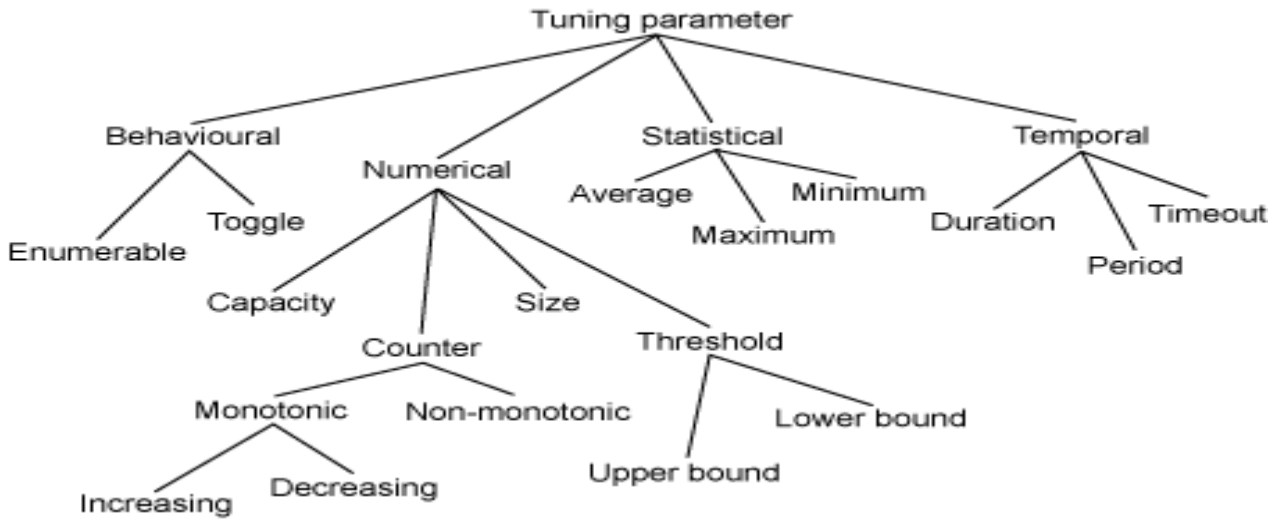

Fig. 2: Pattern Taxonomy

Proposed System

1. The proposed approach is used to improve the accuracy of evaluating term weights.
2. The discovered patterns are more specific than whole documents or files.

3. Pattern mining techniques can be used to find various text patterns.

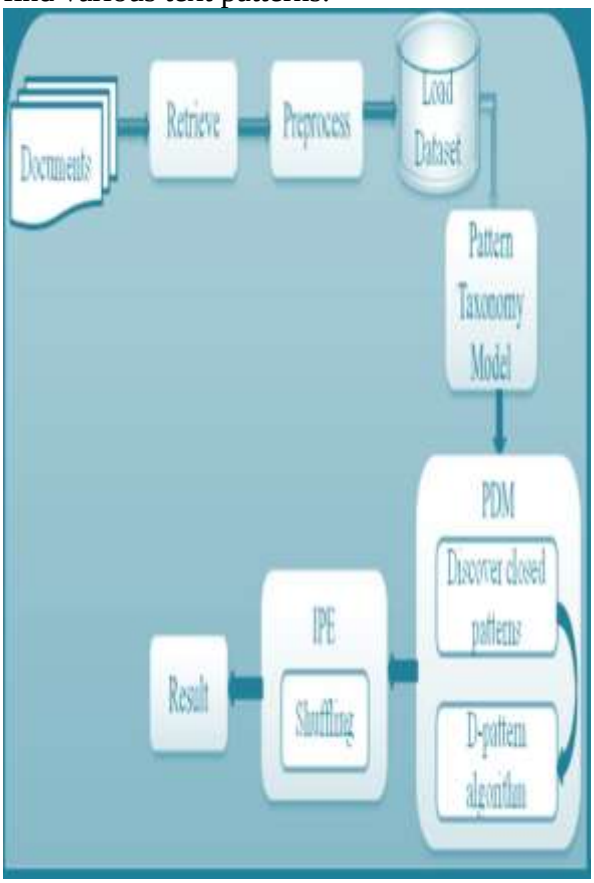

Figure contains the following blocks:

1. Loading Document: In this module, to load the list of all documents. The user to retrieve one of the documents. This document is given to next process. That process is preprocessing.

2. Text Preprocessing: The retrieved document preprocessing is done in module. There are two types of process is done.1) stop words removal 2) text stemming. Stop words are words which are filtered out prior to, or after, processing of natural language data. Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

3. Pattern Taxonomy Process: In this module, the documents are split into paragraphs. Each paragraph is considered to be each document. In each document, the set of terms are extracted. The terms, which can be extracted from set of positive documents.

4. Pattern Deploying: The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated.

5. Pattern evolving : In this module used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. If partial conflict offender contains in positive documents, the reshuffle process is applied.

Experimental Data Set

The most popular used data set currently is RCV1, which includes 936,435 news articles for the period between 20 August 2000 and 19 August 2014. These documents were formatted by using a structured XML schema. TREC filtering track has developed and provided two groups of topics (100 in total) for RCV1 [37]. The first group includes 70 topics that were composed by human assessors and the second group also includes 70 topics that were constructed artificially from intersections topics. Each topic divided documents into two parts: the training set and the testing set. The training set has a total amount of 7,145 articles and the testing set contains 78,344 articles. Documents in both sets are assigned either positive or negative, where “positive” means the document is relevant to the assigned topic; otherwise “negative” will be shown. All experimental models use “title” and “text” of

XML documents only. The content in "title" is viewed as a
paragraph as the one in "text" which consists of paragraphs. For dimensionality reduction, stopword removal is applied and the Porter algorithm [33] is selected for suffix stripping. Terms with term frequency equaling to one are discarded.

Measures

Several standard measures based on precision and recall are used. The precision is the fraction of retrieved documents that are relevant to the topic, and the recall is the fraction of relevant documents that have been retrieved. The precision of first K returned documents top-K is also adopted in this paper. The value of K we use in the experiments is 20. In addition, the breakeven point (b=p) is used to provide another measurement for performance evaluation. It indicates the point where the value of precision equals to the value of recall for a topic. The higher the figure of b=p, the more effective the system is. The b=p measure has been frequently used in common information retrieval evaluations. In order to assess the effect involving both precision and recall, another criterion that can be used for experimental evaluation is F-measure [20], which combines precision and recall and can be defined by the following equation:

Reference

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

[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] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "WordSequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

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

[11] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[12] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning (ICML '98),, pp. 137-142, 1998.

[13] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.

[14] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans.

Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
[15] 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.
[16] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.
[17] D.D. Lewis, "Evaluating and Optimizing Automous Text Classification Systems," Proc. 18th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '95), pp. 246-254, 1995.