Comparative Study and Analysis of Mining Methods in Text Mining

Kavya Jain[1] Dr. Siddharth Choubey[2]
[1,2]Department of Computer Science and Engineering
[1,2]Shri Shankaracharya Technical Campus, Junwani, Bhilai

Abstract— In the text document, huge data mining techniques have been used in order to mine useful pattern. Text mining can be used to haul out the large catalog or datasets from the document or paragraph. The text mining technique is used on the existing term-based technique and produces the difficulty of polysemy and synonymy. The frequent pattern based loom performs better than the term-based ones, but sometime this experiment does not work. This paper includes the procedure of pattern deploying and pattern evolving and enhances the effectiveness of discovering patterns for finding out useful information.

Key words: Data Mining; Text Mining; Frequent Pattern Mining; Term-Based Method; Pattern Evolution

I. INTRODUCTION

Text mining is a technique that is used to find useful information from large amount of data set. Data mining has rule called as frequent pattern and association rule that is important for finding frequent patterns. The apriority based algorithm and tree structure-based algorithms are used in frequent pattern mining. We are using a tree structure-based algorithms, this algorithm follows a test approach a test support frequency only. Examples are FP tree and FR develop tree.

In the last decade, data mining has been proposed different knowledge tasks. These tasks include sequential pattern mining, maximum pattern mining, association rule and closed pattern mining. The synonymy and polysemy methods are creating a problem in term-based method.

A word that shares the same meaning in other words that is called synonymy and a word that has 2 or more meaning that is called polysemy. This paper proposes a temporal text mining approach for frequent patterns mining. Temporal text mining combines data mining techniques and extracting information upon texting repository. The sequences of events from the sets of documents are extracted in order to track the past events effectively.

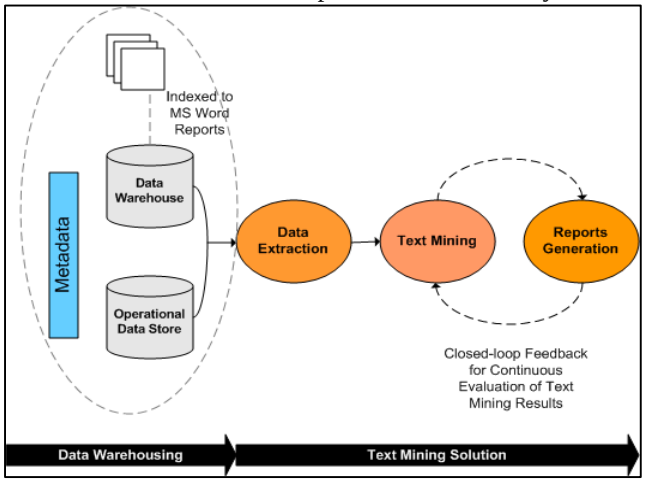

Fig. 1: Text Mining Process

Associated with the given document set are constructed the optimal decomposition of the time periods. The notion of the compressed level decomposition is introduced where each sub-interval consists of consecutive time points having identical information content [3]. Several documents are defined based on the information computed as document set are combined.

Text mining is the discovery of important knowledge in text mining. It is a challenging issue to find Knowledge to help users to find what one wants. Information Retrieval (IR) provided various term-based methods to solve this problem.

There are two main issues of pattern-based method: low frequencies and misinterpretation (miss understanding) [12]. A highly frequent pattern is usually a specific pattern of low frequencies. Many noisy patterns are discovered, if we decrease the minimum support. Misinterpretation means the measure used in patterns mining (e.g., "support" and "confidence") turn out to be not suitable in using discover pattern to answers what users want. In text document, the difficult problem is how to use discovered patterns to accurately evaluate the weights of useful knowledge.

The FP growth algorithm has three advantages: First, it scans database two time and decrease computational cost. Second, generation of candidate item-sets is removed in the FP tree algorithm. Third, it decreased the search space by using the divide and conquers method. The FP growth algorithm has a disadvantage. It cannot have used in incremental mining because when new transactions are added to the database, FP tree requires updating in the data set and complete process are repeated after this. This problem has received attention from researchers in data mining and information retrieval (IR) community.

Many applications such as battlefield surveillance and environment monitoring etc, but the resource-constrained restrictions make it essential for these sensor nodes to conserve energy to increases life-time of the WSN. An early deployment aim was to use these sensors in a passive way for indoor applications. These types of early nodes had the ability to sense scalar data such as temperature, humidity, pressure and location of surrounding objects [Mani M. and Sharma A.K. 2012].

II. PROBLEM IDENTIFICATION

The complexity of recurrent patterns mining from a big quantity of data is generating a huge number of patterns disappointing the smallest amount sustain threshold, especially when min_sup σ is specific low. This is because, all sub-pattern of a recurrent pattern is frequent as well. Consequently, a long pattern contains a numbers of shorter frequent sub patterns. Assorted category of frequent patterns can be excavation from different types of data sets. In this make inquiries, we utilize item sets (sets of item) as a data sets and the wished-for method is for frequent item set mining, that is, the mining of frequent from transactional data set. However, it can be wholesales for additional kinds of frequent patterns.

The problem is more frequently than not decomposed into two sub problems.

1. One is to find those item sets whose occurrence goes beyond a predefined threshold in the databases; those item set are described frequent or large item sets.
2. The second problem is to create involvement rules from those huge item sets with the restriction of minimal self-confidence

A. Intermediate form

Intermediate forms with altering degrees of complexity are fit for different mining purposes. For a fine-grain domain-specific knowledge unearthing task, it is obligatory to perform semantic investigation to derive a amply rich representation to detain the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally costly and often work in the order of a hardly any language per second. It ruins a challenge to see how semantic scrutiny can be made much additional efficient and scalable for very big text corpora.

B. Multilingual Text Refining

Whereas data mining is principally language self-regulating, text mining involves a important language component. It is indispensable to develop text humanizing algorithms that procedure multilingual text documents and create language-independent in-between forms. While for the most part text mining tools focus on dispensation English documents, mining from documents in former languages allows access to beforehand untapped information and offers a novel host of opportunities.

C. Domain Knowledge Integration

Domain knowledge, not catered for by any present text mining equipment, could play a significant role in text mining. Specially, domain information can be second-hand as near the beginning as in the manuscript refining stage. It is attractive to explore how one can obtain advantage of area information to progress parsing efficiency and obtain a more compact intermediate form. Domain knowledge possibly will also play a fraction in knowledge distillation. In a classification or prognostic modelling task, domain knowledge helps to recover learning/mining efficiency as well as the excellence of the learned model (or mined knowledge).It is also interesting to explores how a user's knowledge can be used to initialize a knowledge arrangement and make the discovered knowledge more interpretable.

D. Personalized Autonomous Mining

Current text mining merchandise and applications are still tools considered for trained knowledge specialists. Opportunity text mining tools, as part of the information management systems, should be willingly usable by technological users as well as organization executives. There have been some labors in developing system that understand natural language queries and mechanically perform the suitable mining operation. Text mining tool could also appear in the form of clever personal assistants. Under the agent paradigm, an individual miner would learn a user's outline, conduct text mining operations mechanically, and forward information without requiring a plain request from the user

III. Methodology

Overall method of text mining is depicted in the figure 1[9].Text mining process consists of text pre-processing, texts transformation, features selection, pattern discovery and evaluation as shown in the figure 1.

A. Text Preprocessing

Text preprocesses is the initial step of text mining which reads one text document at time and process it. This step divides into following major three subtasks-

1) Tokenization

Generally text document contain multiples sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc.

2) Stop Word Removing

This process removes stop words such as "a", "are", "the" or any tags like HTML tag etc.

3) Stemming

Stemming is functional after stop word exclusion by reducing the word to its root word. For example. "playing", "played" are stemmed to "play".

B. Text Transformation

Text transformation has the role of adaptation of the text document into words in order to make it will useful for additional processing.

C. Feature Selection

It performs exclusion of the features that are considered unrelated for mining purpose.

D. Pattern Discovery

A pattern discovery is one of the significant processes that use methods to discover pattern. Method includes clustering, classification, summarization, information retrieval, topic extraction etc.

IV. Technique of Text Mining

In the present day languages analysis would work better by the use of computer when compared to human. So the manual techniques were expensive and take more time. To achieve this goal of text mining, there are different recent technology are available by which text mining is performed. In the particular section, different kinds of text mining technologies are discussed for mining texts.

A. Information Extraction

Information extraction is an preliminary step of analyzing shapeless texts. General meaning of this process is simplification of text. It picks out the phrases and finds the associations between them are the key aim of information extraction [10].So that this method is constructive for bulky size of text. To identify phrases, pattern matching approach is used in which comparison of user text with predefined sequence of texts is done. Its extract structured information from unstructured information.

B. Summarization

This process has main aim of specific text from large number of text documents. Manually it is not possible to summarize large document [9].In several research centers, it is not possible to read all text documents that means researcher has no time to read all this documents. They

summarize document and makes summary of document from main points. Summarization has generally two methods that are extractive and abstractive. In extractive method, significant sentences, paragraph etc. are select from a document and then they are joint to forms a small version of texts. Importance of sentence or paragraphs is decided on the base of statistical and linguistic features of information. In abstractive method, complete concept of document is to be expressed in natural language by understanding the whole document. It uses linguistic method to describe the whole documents and forms a new text that delivers significance of an original document.

C. Topic Tracking Basic thought of topic tracking mechanisms are to maintain user profile based on previously searches and guesses other document extremely effectively based on user profile.

Previously searched records are maintained in user profile. This mechanism is useful for the study of new and forthcoming news associated to search. It has one limitation related to search data because it searched relevant data as well as redundant data also. Topic tracking is used in many areas such as radio, news broadcast etc. In the industries, this technology is helpful for examination of the news as compared with its competitor's goods or updates in the market.

D. Classification

It is process of finding main subject of document by totaling metadata and analyzing documents [4]. This method finds count of words and from that count signifies topic of the document. In this procedure, text document is classified into predefined class label. Classification, which is utilised in customer feedback, filtering emails etc is very useful.

E. Clustering

Clustering has no predefined class label, instead of that it uses similarity measures among different object and place similar object in one class and dissimilar objects in another different class. This technique divides text into one group and in that way makes cluster of group. It is a technique of grouping similar documents but varies from categorization. Words are separated very fast then weights are assigned to each word. After calculating similarity, clustering algorithms are applied to generate list of classes.

V. Conclusion and Future Work

Text mining technique is essentially used for extracting pattern from unstructured data. Various techniques for efficiently performing text mining are converse in this paper. So in this paper, our focus is basically on how text is to be mined. We have also converse process of text mining, its application, merit and demerit.

References

[1] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg.
[2] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.
[3] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining," ", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.
[4] Vishal gupta and Gurpreet S. Lehal , "A survey of text mining techniques and applications", journal of emerging technologies in web intelligence, 2009,pp.60-76.
[5] Nitin Jindal and Bing Liu, "Identifying Comparative Sentences in Text Documents", University of Illinois at Chicago.
[6] http://www.cis.upenn.edu/~ungar/KDD/text-mining.html
[7] Mrs.K. Mythili, and Mrs. K. Yasodha, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining", International Journal of Science and Applied Information Technology, Volume 1, No.3, July – August 2012.
[8] Dion H. Goh and Rebecca P. Ang," An introduction to association rule mining: An application in counselling and help seeking behaviour of adolescents", Journal of Behaviour Research Methods39 (2), Singapore, 259-266,2007.
[9] Deepshikha Patel, Monika Bhatnagar, "Mobile SMS Classification", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 (Online), Volume-I, Issue-I, March 2011.
[10] http://people.ischool.berkeley.edu/~hearst/text-mining.html
[11] Ranveer Kaur, Shruti Aggarwal, "Techniques for Mining Text Documents", International Journal of Computer Applications (0975 – 8887) Volume 66– No.18, March 2013.
[12] N. Kanya and S. Geetha ,"Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118.
[13] Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009
[14] Falguni N. Patel, Neha R. Soni," Text mining: A Brief survey", International Journal of Advanced Computer Research (ISSN (pri nt): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.

