www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 5 May, 2014 Page No. 5621-5625

A survey on text mining techniques

Mr. Rahul Patel[#1]***, Mr. Gaurav Sharma***[*2]

[#1]*PG Research Scholar Computer Science Department, Medicaps Institute of Technology & Management,Indore*
[#2]*Assistant Professor Computer Science Department, Medicaps Institute of Technology & Management,Indore*

[1]rahulpatel.cs.svits@gmail.com
[2]er.gaurav622@gmail.com

***Abstract*—text mining is a technique to find meaningful patterns from the available text documents. The pattern discovery from the text and document organization of document is a well-known problem in data mining. Analysis of text content and categorization of the documents is a complex task of data mining. In order to find an efficient and effective technique for text categorization, various techniques of text categorization and classification is recently developed. Some of them are supervised and some of them unsupervised manner of document arrangement. This presented paper discusses different method of text categorization and cluster analysis for text documents. In addition of that a new text mining technique is proposed for future implementation.**

***Keywords*— text mining, classification, cluster analysis, survey**

I. INTRODUCTION

Now in these days, due to computational automation various different text document sources are available. Extraction of patterns and arranging the text document is a key goal of text mining technique development. Text mining is related to data mining, except that data mining tools are considered to handle structured data, but text mining can work with formless or semi-structured data sets. The application of text mining is very popular in emails analysis, digital libraries and others

The text mining techniques starts with collection of text documents (text repository), than a text mining tool for pre-processing is applied. The pre-processing technique clean and format the data, additionally that is responsible for extracting the meaningful features from these documents. In next step the text mining techniques such as clustering or classification algorithm is taken place to arrange the documents. Figure 1 can describe the whole process of text mining [1].

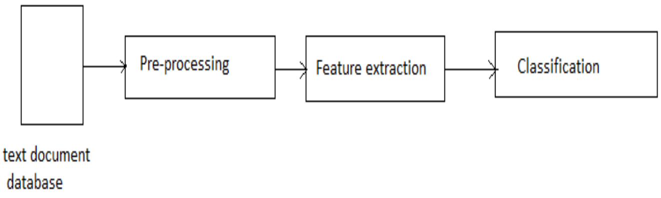

Figure 1 text mining process

This section of the paper provides the general overview of text mining and their applications. In the next section describes the recent research and development over the text mining techniques.

II. RECENT STUDIES

This section of the paper explores recent efforts and contributions on text mining techniques. Therefore a number of research article and research papers and their contributions are placed in this section.

Many data mining techniques have been planned for mining valuable patterns in text documents. However, how to successfully use and update exposed 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 troubles of polysemy and synonymy.This paper presents an inventive and valuable pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to advance the effectiveness of using and updating discovered patterns for finding appropriate and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance [2].

The "helpfulness" characteristic of online user reviews helps consumers deal with information overloads and facilitates decision-making. However, many online user reviews require sufficient helpfulness votes for other users to assess their true helpfulness level. Text mining techniques are employed to remove semantic characteristics from review texts.Our findings also advise that reviews with strong opinions receive more kindness votes than those with mixed or neutral opinions. This paper sheds light on the considerate of online users' helpfulness voting activities and the design of a enhanced helpfulness voting mechanism for online user review systems [3].

Knowledge bases and controlled summaries are playing a important part in many applications, such as text summarization, question answering, essay grading, and semantic search. Although, many systems (e.g., DBpedia and YaGo2) offer vast knowledge bases of such summaries, they all suffer from incompleteness, inconsistencies, and wrongness. These troubles can be addressed and much enhanced by combining and integrating different knowledge bases, but their large sizes and their reliance on diverse terminologies and ontologies make the task very difficult. In this demo, we will exhibit a system that is achieving good success on this task by: i) employing available interlinks in the current knowledge bases (e.g. external Link and redirect links in DBpedia) to combine information on individual entities, and ii) using widely available text corpora (e.g. Wikipedia) and our IBminer text-mining system, to produce and verify structured information, and settle terminologies across different knowledge bases. We will also express two tools designed to bear the integration process in close collaboration with IBminer. The first is the InfoBox Knowledge-Base Browser (IBKB) which offers structured summaries and their provenance, and the second is the InfoBox Editor (IBE), which is designed to advise relevant attributes for a userspecified subject, whereby the user can easily improve the knowledge base without requiring any knowledge about the internal terminology of individual systems [4].

Analyzing large textual collections has develop increasingly challenging given the size of the data existing and the rate that more data is being created. Topic-based text summarization methods coupled with cooperative visualizations have offered promising approaches to address the challenge of evaluating large text corpora. As the text corpora and vocabulary grow larger, more topics require to be created in instruction to capture the significant latent themes and nuances in the corpora. However, it is tough for most of recent topic-based visualizations to represent large number of topics without being jumbled or illegible. To enable the representation and navigation of a large number of topics, we offer a visual analytics system - HierarchicalTopic (HT).User interactions are delivered for users to make variations to the topic hierarchy based on their mental model of the topic space We have also directed a user study to quantitatively calculate the effect of hierarchical topic structure. The study results disclose that the HT leads to quicker identification of huge number of related topics. We have solicited user feedback during the tests and incorporated some suggestions into the current version of HierarchicalTopics [5].

Recent studies have exposed that emerging modern machine learning techniques are beneficial to statistical models for text arrangement, such as SVM. In this study, we converse the applications of the support vector machine with combination of kernel (SVM-MK) to enterprise a text classification system. Conflicting from the average SVM, the SVM-MK uses the 1-norm based object utility and accepts the convex combinations of single trait basic kernels. Only a linear programming problem desires to be resolved and it significantly reduces the computational costs. More imperative it is a transparent model and the optimal characteristic subset can be obtained automatically. A genuine Chinese corpus from Fudan University is used to express the good performance of the SVM-MK [6].

This section describes the recent research and contributions on text mining and classification methodologies. In the next section some popular approaches to detect informative patterns are discussed in next section.

III. TEXT MINING TECHNIQUES

There are different kinds of techniques available by which the text pattern analysis and mining is performed. Some of the essential techniques are discussed in this section.

A. *Information Extraction*

A starting point for computers to examine unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. The software infers the relations between all the identified people, places, and time to deliver the user with significant information. This technology can be very helpful when dealing with large volumes of text. Traditional data mining assumes that the information to be “mined” is previously in the form of a relational database. Unfortunately, for many applications, electronic information is only obtainable in the form of free natural language documents rather than structured databases. Since IE addresses the difficulty of transforming a corpus of textual documents into a extra structured database, the database constructed by an IE module can be provided to the KDD module for advance mining of knowledge as illustrated in Figure 2.

B. *Topic Tracking*

A topic tracking system mechanism by custody of user profiles and, based on the documents the user views, guess other documents of interest to the user. Yahoo offers (www.alerts.yahoo.com) free topic tracking tool that permits users to choose keywords and informs them when news relating to those topics becomes existing. Topic tracking methodology have its own limitations, however. For example, if a user sets up an alert for "text mining", s/he will receive numerous news stories on mining for minerals, and very few that are really on text mining. Some of the improved text mining tools let users select specific categories of interest or the software routinely can even infer the user's concern based on his/her reading history and click-through information.

C. *Summarization*

Text summarisation is enormously helpful for trying to figure out whether or not a extensive document meets the user's needs and is worth reading for advance information. With huge texts, text summarization software procedures and précises the document in the time it may take the user to read the first paragraph. The key to summarisation is to decrease the extent and feature of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to recognize people, places, and time, it is still complex to teach software to analyze semantics and to interpret meaning.

D. *Categorization*

Categorization engage identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often delight the document as a "bag of words."Rather, categorization only calculate words that emerge and, from the counts, identifies the main topics that the document covers. Categorization often relies on a vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, synonyms, and related terms. Categorization utensils normally have a technique for grade the documents in order of which documents have the most content on a specific topic.

E. *Clustering*

Clustering [7] is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results. A basic clustering algorithm generates a vector of topics for each document and determines the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents.

F. *Concept Linkage*

Concept linkage tools [3] attach related documents by identifying their commonly-shared idea and help users find information that they perhaps wouldn't have establish using conventional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

G. *Information Visualization*

Visual text mining, or information visualization [3], puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. DocMiner as shown in figure12, is a tool that shows mappings of large amounts of text, allowing the user to visually analyze the content. The user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them, with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own.

H. *Question Answering*

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question. Many websites that are equipped with question answering technology, allow end users to “ask” the computer a question and be given an answer. Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

I. *Association Rule Mining*

Association rule mining (ARM) [33] is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. ARM refers to the discovery of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently occur. Similar to the idea of correlation analysis (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may contain two or more variables.

This section provides the overview of text mining techniques and methodologies by which suitably text data becomes classifiable in next we discuss the data mining algorithms that are frequently consumed in the text mining and classification tasks.

IV. TEXT MINING ALGORITHMS

There are various algorithms of data mining is available for efficient classification and categorization. The discussion about whole methods and technique are not much feasible here therefore a little overview is proving in this section.

A. *k nearest neighbour*

In the text mining domain the k nearest neighbour algorithm is a classical and frequently used technique. In order to find a query text k nearest neighbour classifier is outperforms. This method estimates the distance between two strings for comparison and classify the text on the basis of distance.

$$d_A(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

Where x and y represents the data instances and d is distance between x and y. The main advantage of this algorithm is high accurate classification. On the other hand the major disadvantage is resources consumption such as memory and time.

B. *Support vector machine*

This approach is a one of most efficient and accurate classification algorithm. In this approach concept using hyper-plans and dimension estimation based technique are used to discover or classify the data. The main advantage of this algorithm is to achieve high accurate classification results. But that is quite complex to implement.

C. *Bayesian classifier*

That a probability based classification technique that is uses the word probability to classify the text data. In this classification scheme based on previous text and patterns data is evaluated and the class possibility is measured. That is some time slow learning classifier additionally that do not produces the more accurate results.

D. K-mean clustering

This technique is also a classical approach of text categorization. That uses the distance function as k nearest neighbour classifier to cluster data. That is an efficient method of text mining in order to preserve the resources, but accuracy of this cluster approach is susceptible due to initial cluster center selection process. In addition of that hierarchical schemes of text categorization is available which are not much efficient for cluster formation or categorization but comparative accuracy is much reliable than k-mean clustering.

V. CONCLUSIONS

In this paper various techniques and methods are discussed for efficient and accurate text mining. In addition of that the efficient algorithms are also learned. Due to observation a promising approach is obtained given in [5]. According to the analyzed methods an improvement over the [5] is suggested. In near future the proposed technique is implemented using JAVA technology and the comparative results are provided.

ACKNOWLEDGMENT

I wish to thanks who directly and indirectly contribute in paper, First and foremost, I would like to thank Prof. Gaurav Sharma for his most support and encouragement. He kindly read my paper and offered valuable detailes and provide guidlines. Second,I would like to thanks all the authors whose paper i refer for there direct and indirect support to complete my work.

REFERENCES

[1] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[2] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering. C Copyright 2010 IEEE

[3] Qing Cao, Wenjing Duan, Qiwei Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach', 0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009

[4] Hamid Mousavi, Shi Gao, Carlo Zaniolo, "IBminer: A Text Mining Tool for Constructing and Populating InfoBox Databases and Knowledge Bases", Proceedings of the VLDB Endowment, Vol. 6, No. 12, Copyright 2013 VLDB Endowment 21508097/13/10...$ 10.00.

[5] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky, "HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 19, NO. 12, DECEMBER 2013

[6] Liwei Wei, Bo Wei, Bin Wang, "Text Classification Using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58, doi:10.4236/jsea.2012.512b012 Published Online December 2012