Applications and Techniques on Text Mining

Bhuvaneshwari. L[1] O. Vidhya[2]
[1]M. Sc. Student [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Dr. SNS Rajakalshmi College of Arts and Science, Coimbatore India

Abstract— In text document, huge data mining techniques have been used for mining useful pattern. Text mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. The pattern discovery from the text and document establishment of document is a well-known problem in data mining. Which is used to extract interesting information or intelligence from the text documents. In addition of that a new text mining technique is prospective for future implementation.

Key words: Text Mining Framework, Classification, Clustering, Survey

I. INTRODUCTION

The text excavating techniques starts with collection of text documents, than a text mining tool for pre-processing is applied. Text mining is a technique that is used to find useful information from large amount of data sets. This paper proposes a temporal text mining approach for frequent pattern mining. The service might be essentially a document classification tool which retrieves documents in response to a string of keywords. The intermediate form of entry representation mining depend specific domain. This is a vital step, of knowledge discovery process. The pre-processing technique clean and format the data, additionally that is responsible for extracting the evocative features from these brochures. The notion of compressed level decomposition is introduced where each subinterval consists of consecutive time points having identical information content. This argues strongly in favour of an underlying architecture which rigging loosely coupled, interchangeable processors, several documents are defined based on the statistics computed as brochures sets are combined. In next step the text mining techniques such as clustering or classification design is taken place to arrange the brochures.

II. TEXT MINING FRAMEWORK

Text mining is the process of extracting interesting information or education or more from the unstructured text. As the text is in shapeless form. It is quite difficult to deal with it. Finding –nuggestel of interesting information from the usual language text is the intention of text mining.

A. Stage-1: Pre-processing Text

Excavating from pre-processed text is easy as compare to natural languages documents. So, pre-processing of documents that are from some another sources is an important task during text mining process before applying any text mining technique.

B. Stage-2: Text mining technique is used:

This is an important stage which the selected algorithm is applied on text in order to process the text. The algorithm such as clustering, classification, summarization, inteligence extractions or visualizations which are explained next could be used.

C. Stage-3: Analysis of text:

Here the outputs are analyzed for discovering the knowledge. Numerous tools such as link discovery tool can be used or the outputs can be visualized so that users could navigate through in order to achieve the perspective.

III. TEXT MINING TECHNIQUES

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 levels of text. Backward looking data mining assumes that the information to be mined is previously in the form of a relational database.

B. Topic Tracking

A topic tracking system mechanism by custody of user profiles and, based on the brouchers the user views, guess other documents of interest to the user. 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 summarization is enormously useful for proving to figure out whether or not an extensive document meets the user's needs and is worth reading for advance information. With huge texts, text summarization software procedures and precises the document in the time it may take the user to read the first paragraph. The key to summarization is to decrease the extent and feature of a document while retaining its main points and overall meaning.

D. Categorization

Categorization engage recognition the main themes of a document by placing the document into a pre-defined set of topics. When cordination a document, a computer program will often delight the document as a –bag of words. Instead, categorization only calculate words that arrive and, from the counts, identifies the main topics that the document covers. Categorization most often relies on a vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, substitutes, and related condition. Categorization utensils usualy have a method for grade the documents in direction of which documents have the most content on a specific topic.

E. Clustering

Clustering is a technique used to group similar brochures, but it differs from categorization in that certificate are clustered on the fly instead of the use of predefined contents. Another advantage of clustering is the documents can emerge in multiple subtopics, thus ensuring that a useful brochures will not be absent from search results. A basic clustering procedures generates a vector of topics for each document and determines the weights of how well the document fits into each cluster.

F. Concept Linkage

Concept linkage tools attach related documents by identifying frequently shared idea and help users find information that perhaps wouldn't have establish using conventional searching procedures. Concept linkage is a valuable idea in text mining, espectionaly in the biomedical fields where so much study has been done that it is no chance for researchers to read all the machinery and make organizations to inverse research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot.

G. Information visualization

Graphic text mining, or materials visualization, puts large textual literature in a visual hierarchy or map and provides browsing capabilities. The government can use information visualization to fingerprint terrorist networks or to find information about offenses that may have been previously deliberation unconnected.

IV. Application of Text Mining

A. Digital Libraries

Numerous text mining procedures and tools are in use to as certain the patterns and trends from journals and proceedings from humongous amount of repositories. These sources of information help in the field of research and development. Libraries are a great source of information for the researchers and digital libraries are endeavoring to the significance of their collection. It provides a novel method of organizing information in such a way that make it possible to available trillions of documents online. In text mining process various operation are performed like documents choosing, improvement, extracting information and interfrence materials among the documents and producing instinctive co-referencing and summarization.

B. Academic and Research Field

In education field, various text mining tools and techniques are used to analyze the educational trends in specific region, student's interest in specific field and employment ratio. Use of text mining in research field help to find and classify research papers and relevant material of different fields at one place. The use of k-means clustering and other techniques help to identify the attributes of relevant information. Students performance in different subjects can be accessed and how different attributes effect the selection of subjects.

C. Life Science

Life science and health care industries are generating large amount of textual and numerical data regarding patients record, diseases, medicines, symptoms and treatments of diseases and many more. The medical records contain varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process very difficult.

D. Social Media

Text mining software packages are available for analyzing social media applications to proctor and study the online swamp writing from internet news, blogs, email etc. Text mining tools help to identify and analyze number of posts, likes and followers on the social media network. This kind of analysis show the people reaction on different posts, news and how it spread neck of woods.

E. Business Intelligence

Text mining plays a significant role in business intelligence that help organizations and enterprises to analyze their customers and competitors to take better decisions. It provides a deeper insight about business and give information how to improve the customer satisfaction and gain competitive advantages. The text mining tools like IBM text analysis. Rapid miner, GATE help to take decisions about the organization that generate alerts about good and bad appearance, market change over that help to take remedial actions. It also helps in telecommunication industry, business and commerce applications and customer chain management system.

V. Conclusion

Frequent item set mining, closed pattern mining, sequential pattern mining, association rule mining and closed pattern mining these all techniques are used in data mining frequent pattern techniques. In this paper various techniques and methods are discussed for efficient and accurate text mining. There is already evidence that for som NLP tasks the use of the Grid as opposed to a single system can improve performance time over very large data sets, and the results improve as the data set gets larger. Domain knowledge integration, varying concepts granularity, multilingual text refinement, and natural language processing ambiguity are major issues and challenges that arise during text mining process. In future research work,we will focus to design algorithms which will help to resolve issues presented in this work.

References

[1] Ms. SonamTripathi 1, asst prof. Tripti Sharma2 "A survey paper for finding frequent pattern in text mining"
[2] Mr. Rahul patel#1, mr. Gaurav sharma*2 "A survey on text mining techniques"
[3] Ramzantalib_, Muhammad kashifhanify, shaeel aayeshaz, and fakeeha fatimax "text mining: techniques, application and issues"
[4] Divyanasa "text mining techniques- a survey"
[5] Claire grover harry halpinewankleintjochen L. Leidner "A framework for text mining services"
[6] V. Gupta, G.S. Lehal – "a survey of text mining techniques and applications"-, journal of emerging technologies in web intelligence,2009.
[7] B.L. Narayana and S.P. Kumar, "A new clustering technique on text in sentence for text".

