International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 12 (December 2015), PP.45-49

Filtering of Unstructured Text

Sudersan Behera[1], N.Vinay Kumar[2]
[1]Assistant Professor, Department of Electronics and Computer Engineering,
[2]Assistant Professor Sreenidhi Institute of Science & Technology Yamnampet,
Ghatkesar, Hyderabad-501301

Abstract:- Text mining has become an exciting research field as it tries to discover valuable information from unstructured texts. The unstructured texts which contain vast amount of information cannot simply be used for further processing by computers. Therefore, exact processing methods, algorithms and techniques are vital in order to extract this valuable information which is completed by using text mining. In this paper, we have discussed general idea of text mining and comparison of its techniques. In addition, we briefly discuss a number of text mining applications which are used presently and in future.

Keywords:- Retrieval, Extraction, Categorization, Clustering, Summarization.

I. INTRODUCTION

Text mining has become important research vicinity. A very large number of information stored in different places in unstructured structure. Approximately 80% of the world’s data is in unstructured text [1]. This unstructured text cannot be easily used by computer for more processing. So there is a need for some technique that is useful to extract some precious information from unstructured text. These information are then stored in text database format which contains structured and few unstructured fields. Text can be sited in mails, chats, SMS, newspaper articles, journals, product reviews, and organization records [2]. Almost every one of the institutions, government sectors,

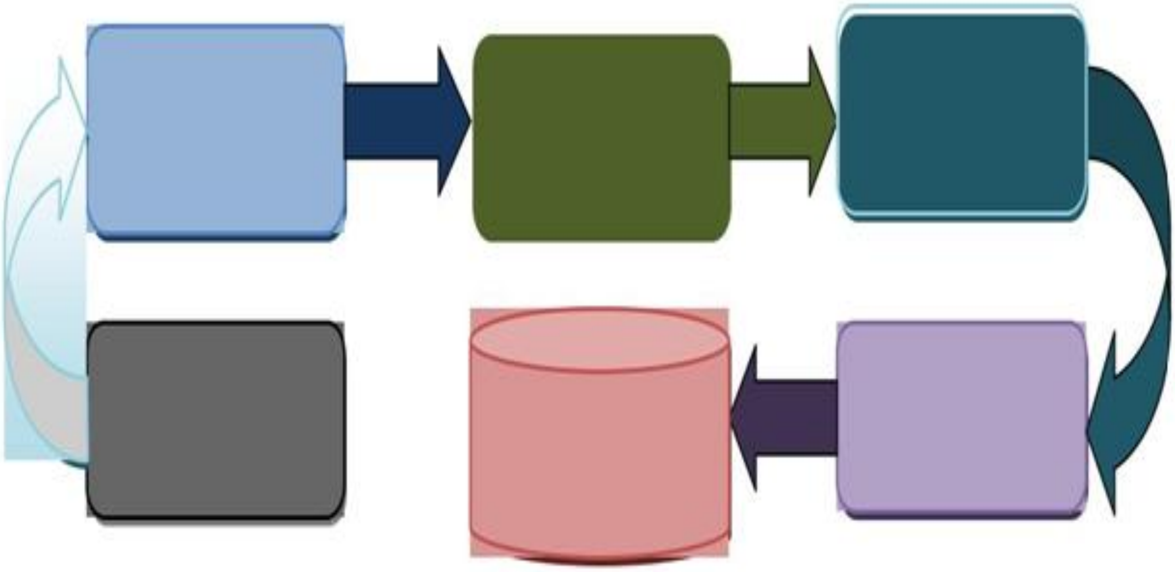

Fig 1: Processing of Text Mining

Organizations and industries information are stored in electronic form.

There are a variety of names for text mining like text data mining, knowledge discovery [4] from textual databases, analysis of intelligent text refers to extracting or retrieve the valuable information from the unstructured text. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases. Text mining discovers new pieces of information from textual data which is earlier unidentified or secret information by extracting it using different techniques. Text mining is a multidisciplinary field, concerning retrieval of information, analysis of text, extraction of information, categorization, clustering, visualization, mining of data, and machine learning.

There are five basic text mining steps as under:

Text mining steps:

a) Collecting information from unstructured data.
b) Convert this information received into structured data
c) Identify the pattern from structured data
d) Analyze the pattern
e) Extract the valuable information and store in the database.

II. BASIC TEXT MINING TECHNOLOGIES

Information Retrieval:

The most well known information retrieval (IR) systems are Google search engines which recognize those documents on the World Wide Web that are associated to a set of given words. It is measured as an extension to document retrieval where the documents that are returned are processed to extract the useful information crucial for the user [3]. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. IR in the broader sense deals with the whole range of information processing, from information retrieval to knowledge retrieval [8]. It is a relatively old research area where first attempts for automatic indexing where made in 1975. It gained increased attention with the grow of the World Wide Web and the need for classy search engines.

Information Extraction:

The goal of information extraction (IE) methods is the extraction of useful information from text. It identifies the extraction of entities, events and relationships from semi-structured or unstructured text. Most useful information such as name of the person, location and organization are extracted without proper understanding of the text [4]. IE is concerned with extraction of semantic information from the text.IE can be described as the construction of a structured image of selected relevant piece information drawn from texts.

Categorization:

Text categorization is a kind of “supervised” learning where the categories are known in advance and firm in progress for each training document. Then, its key projected utilize was for indexing scientific literature by means of controlled words. It was only in the 1990s that the field fully developed with the availability of continuous increasing numbers of text documents in digital form and the requirement to organize them for easier use [5]. Categorization is the assignment of normal language documents to predefined set of topics according to their content. It is a collection of text documents, the process of finding the accurate topic or topics for each document. Nowadays automated text categorization is applied in a variety of contexts from the classical automatic or semiautomatic indexing of texts to personalized commercials delivery, spam filtering, and categorization of Web page under hierarchical catalogues, automatic metadata generation, and detection of text genre, topic tracking and many others [6]. The learning of automated text categorization starts early 1960s. It is a hot topic in machine learning today’s research field.

Clustering:

Clustering is one of the most interesting and important topics in text mining. Its aim is to find intrinsic structures in information, and arrange them into significant subgroups for further study and analysis. It is an unsupervised process through which objects are classified into groups called clusters. The problem is to group the given unlabeled collection into meaningful clusters without any prior information. Any labels associated with objects are obtained solely from the data. For example, document clustering assists in retrieval by creating links between related documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query [8].

Clustering is useful in many application areas such as biology, data mining, pattern recognition, document retrieval, image segmentation, pattern classification, security, business intelligence and Web search. Cluster analysis can be used as a standalone text mining tool to achieve data distribution, or as a pre-processing step for other text mining algorithms operating on the detected clusters.

Summarization:

Text summarization is an old challenge in text mining but in dire need of researcher’s attention in the areas of computational intelligence, machine knowledge and natural language processing. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user. In big organization or company, researcher do not have time to read all documents so they summarize document and highlight summary with main points [4]. A summary is a text that is produced from one or more texts that contains a significant portion of the information, reduced in length and keeps the overall meaning as it is in the original texts. Text summarization involves various methods that employ text categorization, such as neural networks, decision trees, semantic graphs, regression models, fuzzy logic and swarm intelligence. However, all of these methods have a common problem, that is, the quality of the development of classifiers is variable and highly dependent on the type of text being summarized.

III. COMPARISON OF TEXT MINING TECHNIQUES

Text mining uses various numbers of techniques which play an important role. The techniques differ from each other. The information of retrieval technique used unstructured text where it can retrieve valuable information while as the information of extraction extracts the information from structured database. The Summarization technique is used to summarize the document which reduces length and keeps meaning same as it is.

The categorization is supervised process and uses predefined set documents according to their contents. Responsiveness and flexibility of the post-co-ordinate system effectively prohibit the establishment of meaningful relationships because a category is created by individual not the system. While as the clustering is used to find intrinsic structures in information, and arrange them into related subgroups for further study and analysis. It is an unsupervised process through which objects are classified into groups called clusters. Clustering is dealing with high dimensional data, finding interesting pattern associated with data. Another feature is that it is a group of similar type of data and their relationship between them.

Table1: Comparison of text mining techniques

Technique	Characteristics	Tools
Retrieval	**Retrievals valuable information from unstructured text**	**Intelligent Miner, Text Analyst**
Extraction	**Extract information from structured database**	**Text Finder, Clear Forest Text**
Summarization	**Reduce length by keeping its main points and overall meaning as it is**	**Tropic Tracking Tool, Sentence Ext Tool**
Categorization	**Document based categorization**	**Intelligent Miner**
Cluster	**Cluster collection of documents, Clustering, classification and analysis of text document**	**Carrot, Rapid Miner**

IV. APPLICATIONS TEXT MINING

Academic applications

To discover the patterns and trends in the journals and proceedings from huge volume of papers is an essential task in the research field [1]. The matter of importance to publishers who hold large databases of information need indexing for retrieval. This is especially true in scientific disciplines in which highly specific information is often contained within written text. This text mining tool is applied to discover trends on different topics that exist in the proceedings and to show how they change over time. It is also used as topic tracking. Therefore, initiatives have been taken such as Nature's proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD) that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

Bioinformatics

Research work has grown-up in a bioinformatics field, where biomedical literature has become an important research application area for text mining. In the year 2005, the first textbook on biomedical text mining appeared, where it has reported that industry has suggested that 90% of drug targets are derived from the literature. The motivation for this work comes primarily from biologists, who find themselves faced with a massive increase in the number of publications in their field, by keeping up with the related literature is nearly not possible for many scientists [7]. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature in facilitating new innovation in a more efficient manner. One online text mining application in the biomedical literature is that combines biomedical text mining with network visualization as an Internet service. Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that corresponds to instances of concepts that are of interest to biologists. Entity recognition is becoming increasingly important with the massive increase in reported results due to high throughput experimental methods. It can be used in several higher level information access tasks such as relation extraction, summarization and question answering [10].

Copyright and Customer Profile Analysis

The copyright analysis developed to a large application area in recent years because of the increased number of copyright applications. The supervised and unsupervised techniques are applied to analyze copyright documents and to support companies and also the copyright office in some countries to their work. The challenges in copyright analysis consist of the length of the documents, which are larger than documents usually used in text classification, and the large number of available documents in a corpus [6].

Companies use text mining to draw out the occurrences and instances of key terms in large blocks of text such as articles, Web pages, complaint forums. The software converts the unstructured data formats into topic structures and semantic networks which are important information drilling tools. By studying the semantic network, one can learn the general quality of the complaints, reasons for complaining. It also finds common words used in complaints and their relationships to other words in the text via semantic weight [9, 10].

Internet Security

The use of text mining tool in security field has become an important matter. A lot of text mining software packages is marketed for security applications, particularly monitoring and analysis of online plain text sources such as Internet news, blogs, mail etc. for security purposes [7]. It is also involved in the study of text encryption/decryption. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, online chats. Email is used in many legitimate activities such as messages and documents exchange. Unfortunately, it can also be misused, for example in the distribution of unwanted junk mail, mailing offensive or bullying materials. The explosive growth of unsolicited e-mail, more commonly known as spam, over the last years has been undermining constantly the usability of e-mail. One solution is offered by anti-spam filters. Most commercially available filters use black-lists and hand-crafted rules. Since time is crucial and given the scale of the problem, it is infeasible to monitor emails or online chat normally. Thus automatic text mining tools offer a considerable promise in this area [10].

V. CONCLUSION

Text mining generally refers to the process of extracting valuable information from unstructured text. In this survey of text mining, several text mining techniques and its applications in various fields have been discussed. A comparison of different text mining has been shown which can be further enhanced. Text mining algorithms will give us useful and structured data which can reduces time and cost. Hidden information in social network sites, bioinformatics and internet security etc. are identified using text mining is a major challenge in these fields. The advancement of web technologies has lead to a tremendous interest in the classification of text documents containing links or other information.

REFERENCES

[1]. Vallikannu Ramanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.

[2]. Vidya K A, G Aghila, “Text Mining Process, Techniques and Tools: an Overview”,

[3]. International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No 2, pp.613-622.

[4]. R.Sagayam, S.Srinivasan, S.Roshini, “A Survey of Text Mining: Retrieval, Extraction and

[5]. Indexing Techniques”. Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

[6]. Vishal Gupta and Guruprit Lehal, “A Survey of Text Mining Techniques and Applications”,

[7]. Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[8]. Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

[9]. Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.

[10]. Falguni N. Patel, Neha R. Soni,"Text mining: A Brief survey", International Journal of Advanced Computer Research, ISSN (Online):2277-7970, Vol. 2, No. 4, Issue-6, Dec 2012.

[11]. Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319-7242, Vol 3 Issue 5, May 2014, pp.5621-5625

[12]. Seth Grimes, “The developing text mining market”, white paper, Text Mining Summit Alta Plana Corporation, Boston, 2005, pp. 1-12.

[13]. Shaidah Jusoh and Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining", IJCSI, ISSN (Online): 1694-0814, Vol. 9, Issue-6, No. 2, November 2012.