IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 46-51
www.iosrjournals.org

A Survey Paper on Text Mining - Techniques, Applications And Issues

*Mrs.B.Meena Preethi[1], Dr.P.Radha[2]
[1] *Assistant Professor, Department of BCA &M.Sc.SS Sri Krishna Arts and Science College, Kuniamuthur,Coimbatore, India)*
[2]*Assistant Professor, PG & Research Department of Computer Science,Government Arts College,Coimbatore, India)*

Abstract: *Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The recovery of similar patterns and trends to see the text data from huge volume of data is a big issue. Text mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. There lies many techniques and tools to mine the text documents and discover the information for future and process in decision making. The choice of selecting the right and appropriate text mining technique helps to recover the speed and slows the time and effort required to get valuable information.This paper briefly discusses and analyze the text mining techniques and their applications. With the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form. Thus, it has become essential to build better techniques and algorithms to get useful and interesting data from the large amount of textual data.Hence, the field of information extraction and text mining became popular areas of research, to get interesting and needful information.*

Keywords: *Classification; Text Mining Algorithm; Knowledge Discovery; Applications; Information Extraction; Information Retrieval; Patterns*

I. Introduction

Text mining is defined as ―"the separation of hidden and potentially needful information from textual data'' [1]. Text Mining is a new area that searches to extract meaningful data from text language that is natural.It can be designed as the flow of analyzing text to separate information that is needful for a specific purpose. Examining with the type of data stored in databases, text is not designed, ambiguous, and hard to process. However, in today's culture, text is the most commercial way for the formal exchange of information.Text mining deals with texts whose function is the communication of real information or opinions. Text mining is same as data mining, except the data mining tools [2] are designed to use structured data from databases, also text mining can work in fields with unstructured or semi-structured data sets like emails, text documents and HTML files etc. As a result, text mining has a far better solution. Text mining is a process to get interesting and significant patterns to explore knowledge from textual databases [3]. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics [3]. Text mining plays with real language text that are stored in semi-structured and unstructured data format [4]. Text miningtechniques are continuously applied in industry, academia, web applications, internet and other fields [5]. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [6].Text mining is the common process of structuring the input text data(which usually includes the methods like parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and final correction and interpretation of the output.

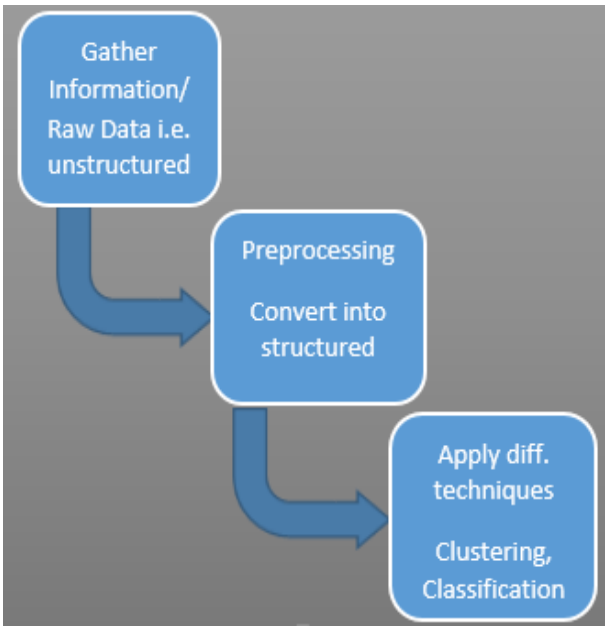

Fig.1: Basic Process of Text Mining

The term ―text mining is most commonly used to relate any system that examines huge quantities of real language text and finds lexical or linguistic usage methods in an attempt to extract needful information.

Areas of Text Mining

Text analysis involves data retrieval, information extraction, data mining techniques includes association and link analysis, visualization and predictive analysis. The aim is, essentially to turn text (unstructured data) into data (structured format) for analysis, via the use of natural language processing (NLP) methods.

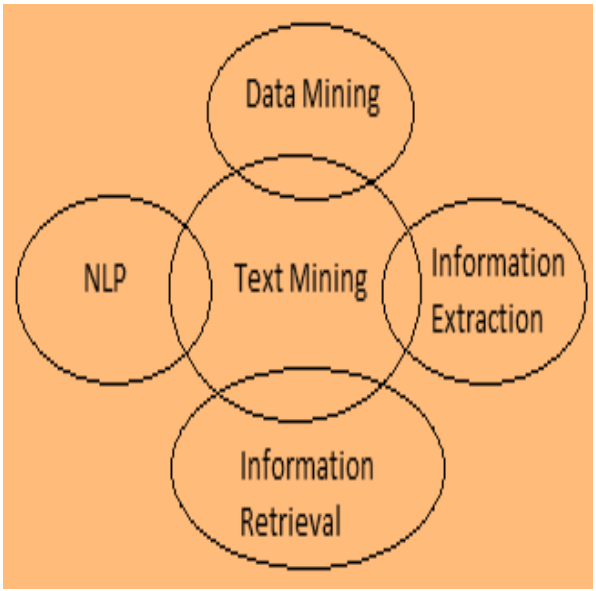

Fig.2: Text mining areas

2.1Information Retrieval (IR)

Information retrieval is designated as full form to document retrieval where the documents are returned and processed to condense or get the particular information retrieved by the user. Thus document retrieval could be followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage using techniques. IR systems helps in to narrow down the set of documents that are relevant to a particular problem. As text mining goes on applying very difficult algorithms to huge document collections, IR can speed up the analysis significantly [8] by decreasing the count of documents for analysis.

2.2 Data Mining (DM)

Data mining can be loosely described as looking for patterns in data. It can be more fully characterized as the extraction of hidden, previously unknown, and useful information [10] from data. Data mining tools can predict behaviors and future trends, allowing businesses to make positive, knowledge based decisions. Data mining tools can answer business questions that have traditionally been too time consuming to resolve. They search databases for hidden and unknown patterns, finding critical information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

2.3 Natural Language Processing (NLP)

NLP is one of the oldest and most challenging problems in the field of artificial intelligence. It is the study of human language so that computers can understand natural languages as humans do [8]. NLP research pursues the vague question of how we understand the meaning of a sentence or a document. What are the indications we use to understand who did what to whom [8], or when something happened, or what is fact and what is supposition or prediction? While words - nouns, verbs, adverbs and adjectives [8] - are the building blocks of meaning, it is their correlation to each other within the structure of a sentence in a document, and within the context of what we already know about the world, that provides the true meaning of a text. The role of NLP in text mining is to deliver the system in the information extraction phase as an input.

2.4 Information Extraction (IE)

Information Extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity includes processing human language texts by means of natural language processing (NLP). The recent activities in multimedia document processing like automatic annotation and mining information out of images/audio/video could be seen as information extraction and the best practical and live example of IE is Google Search Engine. It involves defining the general form of the information that we are interested in as one or more templates, which are used to guide the extraction process. IE systems greatly depend on the data generated by NLP systems.

II. What Is Text Mining?

The Concept

Text mining is a burgeoning new field that tries to extract meaningful information from natural language text [6]. It may be characterized as the process of analyzing text to extract information that is useful for a specific purpose. Compared with the kind of data stored in databases, text is unstructured, ambiguous, and difficult to process. Nevertheless, in modern culture, text is the most communal way for the formal exchange of information. Text mining usually deals with texts whose function is the communication of actual information or opinions, and the stimuli for trying to extract information from such text automatically is compelling—even if success is only partial. Text mining, using manual techniques, was used first during the 1980s [11]. It quickly became apparent that these manual techniques were labor intensive and therefore expensive. It also requires too much time to manually process the already growing quantity of information. Over time there was a huge success in creating programs to automatically process the information, and in the last few years there has been a great progress. The study of text mining concerns the development of various mathematical, statistical, linguistic and pattern-recognition techniques which allow automatic analysis of unstructured information as well as the extraction of high quality and relevant data, and to make the text as a whole better searchable. A text document contains characters which together form words, which can be further combined to generate phrases. These are all syntactic properties that together represent already defined categories, concepts, senses or meanings [11]. Text mining must recognize, extract and use the information. Instead of searching for words, we can search for semantic patterns, and this is therefore searching at a higher level.

Process

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

Text Pre-processing

It involves a series of steps that are shown in the below figure

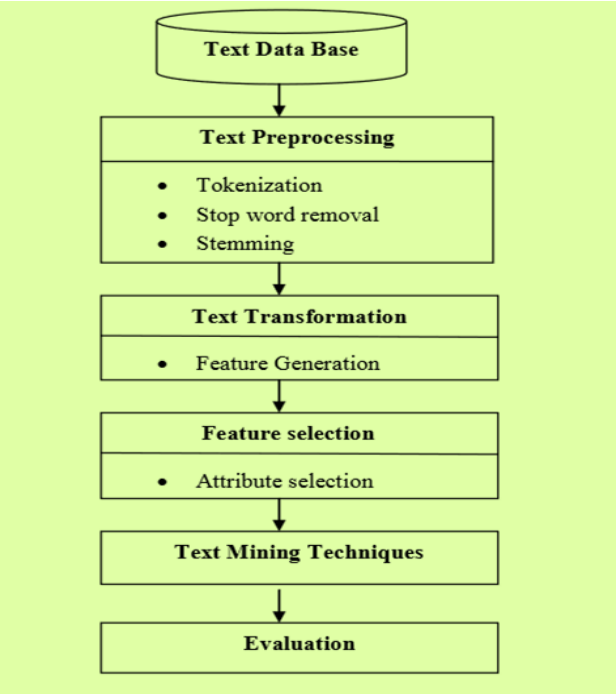

Fig.3: Text preprocessing

3.2.1.1Text Cleanup

Text Cleanup means removing of any unnecessary or unwanted information such as remove ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas.

3.2.1.2Tokenization

Tokenizing is simply achieved by splitting the text on white spaces and at punctuation marks that do not belong to abbreviations identified in the preceding step.

3.2.1.2 Part of Speech Tagging

Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words (OOV problem) and ambiguous word-tag mappings. Rule-based approaches like ENGTWOL [12] operate on a) dictionaries containing word forms together with the associated POS labels and morphological and syntactic features and b) context sensitive rules to choose the appropriate labels during application.

3.2.2Text Transformation (Attribute Generation)

A text document is represented by the words (features) it contains and their occurrences. Two main approaches of document representation are a) Bag of words b) Vector Space.

3.2.3Feature Selection (Attribute Selection)

Feature selection also known as variable selection, is the process of selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the data contain many redundant or irrelevant features. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context. Feature selection technique is a subset of the more general field of feature extraction.

3.2.4Data Mining

At this view the Text mining process is joined with the traditional Data Mining process. Classic Data Mining techniques are used in the structured database that gave the results from the previous stages.

3.2.5Evaluate

Check the result for the correctness, after the corrections the result can be omitted or the generated outcome can be used as an input for the next set of sequence.

III. Application

Text Mining can be applied in many areas [13]. Some of the most common used areas are:

3.1Web Mining

These days web contains a many information about subjects such as persons, companies, products, etc. [14] that may be of huge interest. Web Mining is an important application of data mining techniques to discover hidden and unknown patterns from the Web. Web mining is an important activity of recognizing term implied in large document collection say C, which can be denoted by mapping i.e. C →p [14]. The first take toward any Web-based text mining effort would be to gather a substantial number of web pages having observation of a subject. Then, the question becomes not only to find all the subject developments, but also to separate out those that have the wanted meaning.

3. 2Clustering

Clustering is an unsupervised process to classify the text documents in groups by using different clustering algorithms. In a cluster, same descriptions or designs are grouped got from different documents. Clustering is carried outin top-down and bottom up behavior. In NLP, many typesof mining tools and techniques are used for the determination on unstructured text. Various methods of clustering are distribution, density, centroid, hierarchical and k-mean [22].

4.3Social Media

Text mining software packages are available for analyzing social media applications to monitor and analyze the online plain text from internet news, blogs, email etc. Text mining tools help to identify the number of posts, likes and followers on the social media. This kind of findings show the people reaction on different posts, news and howit gets spread around. It shows the behavior of people belonging to specific age group or communities having similarity and differences in views about the specific post [23], [24].

3.4Resume Filtering

Big companies and headhunters gets thousands and lakhs of resumes from job applicants every day. Obtaining information from resumes with high precision and revising is not an easy task [1]. Instead of constituting a restricted domain, resumes can be written in a multitudinalformats (e.g. structured tables or plain texts), in different languages (e.g. Japanese and English) and in different file types (e.g. Plain Text, PDF, Word etc.). Moreover, writing styles can also be much varied. In the first manual scan of the resume, a recruiter looks for mistakes, educational qualifications, employment history, job titles, frequency of job changes, and other personal information. Exactly getting this information will be the first step in ignoring resumes. Hence, the process of selection of resume is an important task in recruitment.

3.5. Medical and life science

Users frequently exchange information with others about areas of interest or send requests to web-based forums, or ask the expert services [15]. Every people wants to understand particular diseases (what they have), to be told about new therapies, questioned for a second opinion before treatment. Additionally, these forums also indicate seismographs for medical and/or psychological requirements, which are correctly not met by present health care based systems [15].Medias like E-mails, e-consultations, and requests for medical advice through the network have been manually weighed using quantitative or qualitative methods [16]. In order to help the medical experts and to make use of this seismograph function of expert forums, it would be helpful to distinguish visitors' requests instantly. So, particular requests could be directed to the expert or even answered semi-automatically, providing complete monitoring. By creating ―frequently asked questions (FAQs)‖ same alike patient requests [16] and their c] answers could be congregated, even before the particular expert responses. Machine-based conclusions could help the public to handle the mass of information and medical experts to give expert their feedback. An instant classification of amateur requests to medical expert network forums is an heavy task because these requests canbe long and unstructured as an end of mixing, for example, personal experiences with laboratory data. It is a big challenge to find out an correct and important text to take a right decision from an huge biological repository [19]. The records of medicals contain content varying in nature,complex, lengthy and technical vocabulary are used that makethe knowledge discovery process difficult [20].The text mining tools in biomedical field gives an opportunity to obtain valuable information, their association and their relationship among various diseases [21]. Text mining used in biomarker discovery,

pharmaceutical companies, clinical trade analysis, preclinicalsafe toxicity report studies, patent competitive intelligenceand landscaping, mapping of genetical diseases and exploring the specified identifications by using different techniques [18].

IV. Conclusion

The availability of large volume of text based data needed to be analyzed to obtain useful information. Text mining methods are used to find the interesting and important information effectively and efficiently from large amount ofunstructured data. This paper presents a brief overview oftext mining techniques that help to improve the text miningprocess. Specific patterns and sequences are applied and used in orderto obtain and get useful information by ignoring useless details for predictive researches. Selection and use of correct methods and tools according to the domain will help to make the text mining method more easy and efficient. Domain knowledge based integration,concepts based granularity, multilingual text of refinement, and using natural language processing ambiguity are major issues and risks that arise during text mining techniques. Moreover, use of an correct text mining tools in medical field help to calculate the effectiveness of medical treatments that show good effectiveness by comparing different diseases, symptoms . More than any other field, text mining is of great advantage in life science and health care.

References

Journal papers

[1]. Mrs. SayantaniGhosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay. ―A tutorial review on Text Mining Algorithms‖, in International Journal of Advanced Research in Computer and Communication Engineering, volume-1 issue-42012.

[2]. Vishal Gupta, Gurpreet S. Lehal, 2009. ―A Survey of Text Mining Techniques and Applications‖ in Journal of Emerging Technologies in Web Intelligence, Vol. 1 No. 1.

[3]. Shiqun Yin Yuhui Qiu1,ChengwenZhong, 2007. Web Information Extraction and Classification Method .IEEE

[4]. I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library," International Journal on Digital Libraries, vol. 4, no. 1,pp. 56–59, 2004.

[5]. B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," IJSEAT, vol. 3, no. 3, pp. 69–71, 2015.

[6]. C. Ding and H. Peng, "Minimum redundancy feature selection frommicroarray gene expression data," Journal of bioinformatics and computationalbiology, vol. 3, no. 02, pp. 185–205, 2005.

Books

[7]. Navathe, Shamkant B. and ElmasriRamez. ―Data Warehousing and Data Mining‖, in ―Fundamentals of Database Systems‖, Pearson Education pvtInc, (Singapore, 841-872, 2000).

[8]. S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining:predictive methods for analyzing unstructured information. (SpringerScience and Business Media, 2010.)

[9]. Widman LE, Tong DA Arch (Intern Med. 1997), Requests for medical advice from patients and families to health care providers who publish on the World Wide Web. 209-12.

Chapters in Books

[10]. W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of textmining," Communications of the ACM, (vol. 49, no. 9, pp. 76–82, 2006).

[11]. S.-H. Liao, P.-H.Chu, and P.-Y. Hsiao, "Data mining techniques andapplications–a decade review from 2000 to 2011," (Expert Systems withApplications, vol. 39, no. 12, pp. 11 303–11 311, 2012.)

[12]. W. He, "Examining students online interaction in a live video streamingenvironment using data mining and text mining," (Computers in HumanBehavior, vol. 29, no. 1, pp. 90–102, 2013.)

[13]. A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," (Briefings in bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.)

[14]. A. Henriksson, J. Zhao, H. Dalianis, and H. Bostr¨om, "Ensembles of randomized trees using diverse distributed representations of clinical events," (BMC Medical Informatics and Decision Making, vol. 16, no. 2, p. 69, 2016.)

[15]. I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach," (Expert Systems with Applications, vol. 44, pp. 386–399, 2016.)

Thesis

[16]. Daniel Waegel. ―The Development of Text-Mining Tools and Algorithms.Ursinus College, 2006.

[17]. lan H. Witten, ―Text mining‖, University of Waikato, Hamilton, New Zealand

[18]. Johannes C. ScholtesA. Voutilainen.―A syntax-based part of speech analyser‖.In Proc. of the Seventh Conference of the European Chapter of the Association for Computation al Linguistics, pages 157–164, Dublin. Association for Computational Linguistics, 1995

[19]. Johannes C. Scholtes. ―Text-Mining: The next step in search technology‖, DESI-III Workshop Barcelona, 2009.

[20]. Umefjord G, Hamberg K, Malker H, Petersson G FamPract, 2006. The use of an Internet-based Ask the Doctor Service involving family physicians: evaluation by a web survey, 159-66.

[21]. Y. Zhao, "Analysing twitter data with text mining and social networkanalysis," in Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013), 2013, p. 23.