Volume 4, No. 3, March 2013
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICLE
Available Online at www.jgrcs.info

TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS

Lokesh Kumar*[1], Parul Kalra Bhatia[2]
*[1]Department of IT, Amity University, Noida, U.P., India
lokesh.kumar2@student.amity.edu [1]
[2]Department of IT, Amity University, Noida, U.P., India
pkbhatia@amity.edu [2]

Abstract: 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. Text, so it has become essential to develop better techniques and algorithms to extract useful and interesting information from this large amount of textual data. Hence, the area of text mining and information extraction has become popular areas of research, to extract interesting and useful information. This paper, focuses on the concept, process and applications of Text Mining.

Keywords: Text Mining Algorithms, Data Mining, Information Retrieval, Information Extraction,

INTRODUCTION

Text mining is defined as "the non-trivial extraction of hidden, previously unknown, and potentially useful information from (large amount of) textual data'' [1]. Text Mining is a new field that tries to extract meaningful information from natural language text. It can be defined as the process of analyzing text to extract information that is useful for a specific purpose. Compared with the type 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 fascinating - even if success is only partial.

Text mining is similar to data mining, except that data mining tools [2] are designed to handle structured data from databases, but text mining can also work with unstructured or semi-structured data sets such as emails, text documents and HTML files etc. As a result, text mining is a far better solution.

Text mining usually is the process of structuring the input text (usually 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 evaluation and interpretation of the output.

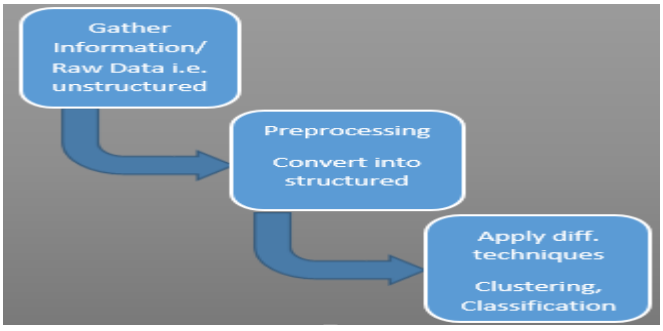

Figure 1: Basic Process of Text Mining

The term "text mining" is commonly used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information.

AREAS OF TEXT MINING

Text analysis involves information retrieval information extraction, data mining techniques including association and link analysis, visualization and predictive analytics [3]. The goal is, essentially to turn text (unstructured data) into data (structured format) for analysis, via the use of natural language processing (NLP) methods.

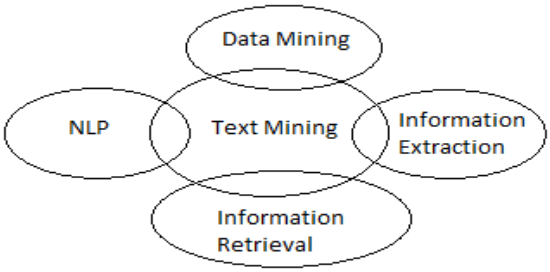

Figure 2: Text mining areas

Information Retrieval (IR):

Information retrieval is regarded as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought 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 involves applying very complex algorithms to large document collections, IR can speed up the analysis significantly [4] by reducing the number of documents for analysis.

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 [4] 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.

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 [5].

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 [5], or when something happened, or what is fact and what is supposition or prediction? While words - nouns, verbs, adverbs and adjectives [5] - 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.

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.

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 [7]. 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 [7]. 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 as shown in figure 3:

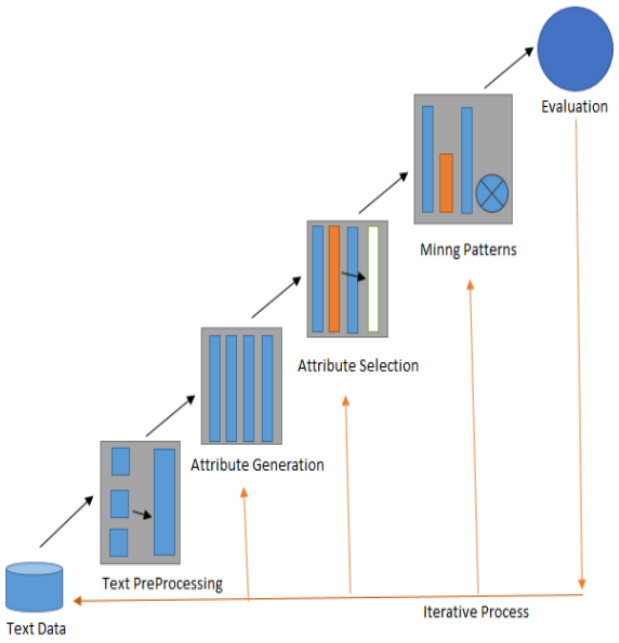

Figure 3. Activities / Process of Text Mining

(a). Text 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.

(b). Tokenization:

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.

(c). 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 [8] 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.

Text 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.

Feature 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.

Data Mining:

At this point the Text mining process merges with the traditional Data Mining process. Classic Data Mining techniques are used in the structured database that resulted from the previous stages.

Evaluate:

Evaluate the result, after evaluation the result can be discarded or the generated result can be used as an input for the next set of sequence.

Applications:

Text Mining can be applied in a variety of areas [9]. Some of the most common areas are:

Web Mining:

These days web contains a treasure of information about subjects such as persons, companies, organizations, products, etc. [10] that may be of wide interest. Web Mining is an application of data mining techniques to discover hidden and unknown patterns from the Web.

Web mining is an activity of identifying term implied in large document collection say C, which can be denoted by a mapping i.e. C →p [10]. The first step toward any Web-based text mining effort would be to gather a substantial number of web pages having mention of a subject. Thus, the challenge becomes not only to find all the subject occurrences, but also to filter out those that have the desired meaning.

Medical:

Users actively exchange information with others about subjects of interest or send requests to web-based expert forums, or so-called "ask the doctor" services [11]. Everyone wants to understand specific diseases (what they have), to be informed about new therapies, ask for a second opinion before one can decide a treatment. In addition, these expert forums also represent seismographs for medical and/or psychological requirements, which are apparently not met by existing health care systems [11].

E-mails, e-consultations, and requests for medical advice via the Internet have been manually analyzed using quantitative or qualitative methods [12]. To help the medical experts and to make full use of the seismograph function of expert forums, it would be helpful to categorize visitors' requests automatically. So, specific requests could be directed to the expert or even answered semi-automatically, thereby providing complete monitoring. By generating "frequently asked questions (FAQs)" similar patient requests [12] and their corresponding answers could be congregated, even before the actual expert responses. Machine-based analyses could help both the public to better handle the mass of information and medical experts to give expert feedback.
An automatic classification of amateur requests to medical expert internet forums is a challenging task because these requests can be very long and unstructured as a result of mixing, for example, personal experiences with laboratory data.

Resume Filtering:

Big enterprises and headhunters receive thousands of resumes from job applicants every day. Extracting information from resumes with high precision and recall is not an easy task [1]. In spite of constituting a restricted domain, resumes can be written in a multitude of formats (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 diversified. In the initial manual scan of the resume, a recruiter looks for mistakes, educational qualifications, buzzwords, employment history, job titles, frequency of job changes, and other personal information [13]. Automatically extracting this information can be the first step in filtering resumes. Hence, automating the process of resume selection is an important task.

SUMMARY AND OUTLOOK

In general Text mining consists of the analysis of text documents by extracting key phrases, concepts, etc. and prepare the text processed for further analyses with data mining techniques. This paper, discussed the concept, process and applications of text mining, which can be applied in multitude areas such as webmining, medical, resume filteration, etc. It also enlighten the hidden potential that lies in the field of text mining and motivated to explore it further.

REFERENCES

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

[2] Navathe, Shamkant B. and Elmasri Ramez. "Data Warehousing and Data Mining", in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, 841-872, 2000.

[3] http://en.wikipedia.org/wiki/Text_analytics

[4] Mrs. Sayantani Ghosh, 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, Vol. 1, Issue 4, 2012.

[5] http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm

[6] Ian H. Witten, "Text mining", University of Waikato, Hamilton, New Zealand

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

[8] 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.

[9] 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.

[10] Shiqun Yin Yuhui Qiu1,Chengwen Zhong, 2007. Web Information Extraction and Classification Method .IEEE

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

[12] 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.

[13] Text Mining Summit Conference Brochure, http://www.textminingnews.com/, 2005

Short Bio Data for the Authors

 **Lokesh Kumar** is pursuing M.Tech in Information Technology from ASET, Amity University. He completed his B.Tech from BBDNITM, Uttar Pradesh Technical University. His area of interests are Data Mining, Information Extraction, Genetic Algorithms, Machine Learning Methods, etc.

 **Parul Kalra Bhatia** is working as an Assistant Professor in the Department of Information Technology, Amity School of Engineering & Technology, Amity University, Noida. She has 8 years experience in the field of Academics and is actively involved in research & development activities. She is pursing Ph.D. (CSE) in the field of Information Retrieval. She has her M.Sc. degree in Computer Science in 2003 and M.Tech in Computer Science and Engineering in 2005 from Banasthali Vidyapith, Rajasthan. Her are of interest includes Text Mining, Information Retrieval- Cognitive and Personalized Copyrights, Intellectual Property Rights, Digital Rights Management, Privacy Rights Management, Advance Database Management System. She has successfully published national and internationalresearch papers.

