International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 2 Issue 7 July 2015

A Survey On : Automated Text Mining

Deepti Gupta[1], Er. Jitendra Dangra[2]
Department of Computer Science, Lakshmi Narain College of Technology
Indore, Madhya Pradesh, India
deepti.gupta003@gmail.com[1], jitendra.dangra@gmail.com[2]

Abstract: *Text Mining has become an important research area. Text mining is the process of extracting useful information or knowledge or patterns from the unstructured text .Automatic text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources. Advantages of automatic text mining approach over the manual and semi-automatic approach are increases effectiveness and savings in terms of expert labour power, and straightforward portability to different domains. In this paper, a Survey of Automatic Text Mining techniques and applications and challenges in text mining have been presented.*

Keywords: *Text Mining, Clustering, Information extraction, Information retrieval, Machine learning, Metadata, Natural language processing etc.*

1. INTRODUCTION

The data stored in the computer can be in any one of the form (i) Structured (ii) Semi Structured and (iii) Unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Text Mining [1] is defined as the process of discovering hidden, useful and important pattern from unstructured text documents [8]. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Text mining is a field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing. Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is generally derived through the designing of patterns and trends through means such as statistical pattern learning. Text Mining is also process of turning text into numeric data, so that it can be used in an analysis or predictive modelling.

Automatic text mining techniques have a long way to go before they rival the ability of people, even without any special domain knowledge, to collect information from large document collections. Automatic text categorization has many applications, for example indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources.

Text mining extracts information from unstructured text. Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modeling of hidden patterns.

The aim of this paper is to give an introduction of automated text mining and overview of text mining system. The paper is organized as follows. Section 2 presents methods of text mining. Section 3 steps involve in text mining Section 4 addressed the challenging issue in text mining. Section 5 presents application of text mining. Section 6 presents proposed work in text mining.

2. METHODS OF TEXT MINING

This section describes the major ways in which text is mined when the input is plain natural language.

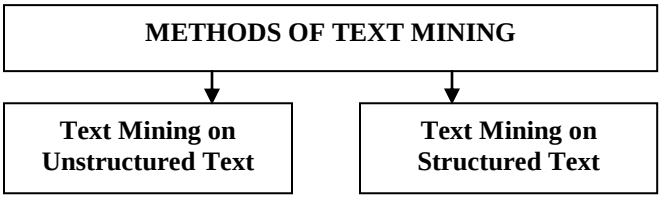

Fig.1: Methods of Text Mining

2.1 Text Mining on Unstructured Text

Here are the various techniques which mine the unstructured text.

2.1.1TextsummarizationA text summarizer produces a summarized representation of its input. It also contains individual documents or groups of documents. Output of text summarization is specific to be human-readable.

2.1.2Document Retrieval Document retrieval is the task of finding and returning the most relevant documents. Traditional libraries provide catalogues that allow users to identify documents based on resources which consist of metadata. Automatic extraction of metadata (e.g. subjects, language, author, key-phrases) is a prime application of text mining techniques. Web search engines are no doubt the most widely-used of document retrieval systems.

2.1.3Information retrieval Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to extract the particular information according to user.

2.1.4Assessing document similarity So many text mining problems involve assessing the similarity between different documents; for example, assigning documents to pre-defined categories and grouping documents into natural clusters.

2.1.5Text categorization Text categorization is the process of categorization of documents to predefined categories according to their content. As in other areas of text mining, until the 1990s text categorization was dominated by ad hoc techniques of "knowledge engineering" that based on categorization rules from human experts and code them into a system that could apply them automatically to new documents. Since then—and particularly in the research community—the dominant approach has been to use techniques of machine learning to infer categories automatically from a training set of pre-classified documents. Many machine learning techniques have been used for text categorization.

2.1.6Document clustering Document clustering is "unsupervised" learning in which there is no predefined category or "class," but groups of documents that belong together are ordered. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query. It is an unsupervised learning technique in which, no pre-defined input-output patterns are there.

2.1.7Language identification Language identification is a particular application of text categorization. A relatively simple categorization task, it provides an important piece of metadata for documents collections. This works particularly well for language identification.

2.1.8 For providing authorship Author metadata is one of the primary attributes of most documents. It is usually known and need not be mined, but in some cases authorship is uncertain and must be guessed from the document text. Authorship ascription is often treated as a text categorization problem. However, there are sensitive statistical tests that can be used instead, based on the fact that each author has a characteristic vocabulary whose size can be estimated statistically from a corpus of their work. Identifying key-phrases In the scientific and technical literature, keywords and key-phrases are attached to documents

2.1.9Extracting structured information An important form of text mining takes the form of a search for structured data inside documents. Ordinary documents are full of structured information: phone numbers, fax numbers, addresses, email addresses, tables of contents, tables, Web addresses etc. Many short documents describe a particular kind of object or event, and in this case elementary structures are combined into a higher-level composite that represent the document's entire content. In constrained situations, the composite structure can be represented as a "template" with slots that are filled by individual pieces of structured information. From a large set of documents describing similar objects or events it may even be possible to infer rules that represent particular patterns of slot-fillers.

2.1.10 Entity extraction Many tasks involve identifying linguistic constructions that stand for objects or "entities" in the world. Often consisting of more than one word, these terms act as single vocabulary items, and many document processing tasks can be significantly improved if they are identified as such. They can aid searching, interlinking and cross-referencing between documents.

2.1.11Information extraction Typical extraction problems address simple relationships among entities, such as finding the predicate structure of a small set of pre-determined propositions[9]. Machine learning has been applied to the information extraction task by using pattern-matching rules that extract fillers for slots in the template.

2.2 Text Mining Structured Text

Here are the various techniques which mine the structured text.

2.2.1Wrapper induction Internet resources that contain relational data—telephone directories, product information, etc.—use formatting mark up to clearly present the information they contain to users. However, with standard HTML, it is quite difficult to extract data from such resources in an automatic way. The XML markup language is designed to overcome these problems by encouraging page authors to mark their content in a way that reflects document structure at a detailed level; but

it is not clear to what extent users will be prepared to share the structure of their documents fully in XML.

2.2.2Document clustering with links Document clustering techniques are normally based on the documents' textual similarity. However, the hyperlink structure of Web documents, encapsulated in the "link graph" in which nodes are Web pages and links are hyperlinks between them, can be used as a different basis for clustering.

2.2.3Determining authority of Web documents The Web's linkage structure is a valuable source of information that reflects the popularity, can be interpreted as importance, authority of Web pages. For each page, a numeric rank is computed. The basic premise is that highly-ranked pages are ones that are cited, or pointed to, by many other pages.

3. STEPS INVOLVED IN TEXT MINING

The steps involved in the process of text mining is shown below

3.1 Text Pre-processing The text pre-processing step is further divided into

3.1.1Tokenization Text documents contain a collection of statements. This step segments the whole text into words by removing blank spaces, commas etc.

3.1.2 Stopword removal This step involves removing of HTML, XML tags from web pages. Then the process of removal of stop words such as 'a', 'is', 'of' etc. is performed.

3.1.3 Stemming Stemming refers to the process of identifying the root of a certain word. There are basically two types of stemming (i) inflectional and (ii) derivational. The most commonly used algorithm is porter's algorithm for stemming.

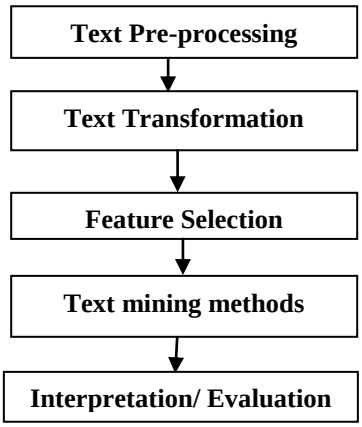

Fig2 : Text Mining Process

3.2Text Transformation Text document is represented by the words it contains and their occurrences. Two approaches used for document representation are

a. Bag of words
b. Vector spaces.

3.3Feature Selection It is also known as variable selection. It is the process of selecting a subset of important features for use in model creation. This phase mainly performs removing features which are redundant or irrelevant. feature selection is the subset of more general field of feature extraction.

3.4Text mining methods At this point Text mining becomes data mining. Data mining methods such as clustering, classification information retrieval etc. can be used for text mining.

3.5 Interpretation/Evaluation Analyzing the results The results generated are used as part of the input for one or more earlier stages.

4. CHALLENGES IN TEXT MINING

The major challenging issue in text mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. Ambiguity means the capability to understand in two or more possible ways. Ambiguity gives a natural language its flexibility and usability so that it cannot be entirely eliminated from the natural language. One word may have multiple meanings. One phrase or sentence can be interpreted in various ways, thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain. On the other hand, most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. According to [10], information extraction system does a more limited task than full text understanding. He pointed that in full text understanding, all the information in the text is presented, whereas in information extraction, the semantic range of the output, the relations will be presented are delimited. However, the growing need for IE application to domains such as functional genomics requires more text understanding. Named entity recognition describes an identification of entities in free text. For example, in biomedical domain, entities would be gene, protein names and drugs. Named entity recognition often forms the starting point in a text mining system, meaning that when the correct entities are recognized, the search for patterns and relations between entities can begin.

Research work in [13] demonstrated on using possibility theory and context knowledge in resolving an ambiguous entity. The obtained results show that the approach was successful; however, the context of the texts should be defined by a user.

5. APPLICATION OF TEXT MINING

There are various applications of Text mining like automatic processing of messages and emails. Some of text mining applications are as follows:

5.1 Patent Analysis Text clustering techniques for patent analysis are often applied to support the analysis of patents in large companies by structuring and visualizing the investigated corpus.

5.2Text Classification for News Agencies In publishing houses a large number of news stories arrive each day. The users like to have these stories tagged with categories and the names of important persons, organizations and places.

5.3Bioinformatics Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that correspond to instances of concepts that are of interest to biologists.

5.4Analysis of the Market trends There is a need, to know the market conditions for the growth of an organization, such as its number of employees, sales, products etc. of the competing organizations. Due to the arrival of Text Mining Techniques, it becomes simple to handle.

5.5Anti-Spam Filtering of Emails 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.

5.6Security applications Many text mining software packages are marketed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs etc. for national security purposes.

5.7Academic applications The issue of text mining is of importance to publishers who hold large databases of information needing indexing for retrieval.

6. PROPOSED WORK

The Text Mining is natural language processing technique for analysis unstructured and structure data for obtaining essential pattern from data. The Text Mining also involves semantic learning processes automatic text classification and categories methodologies that enable the system to automatically recognized the contents and their applications but recently and traditionally developed techniques are not much efficient or accurate in their semantic means in this context a novel approach is required to design by which accuracy and time consumption during automated data classification is improvable technique.

The proposed model incorporate fuzzy logic implementation for distinguish the contents among two given input files and also should be capable to finding similarity among the contents available. In order to simulate presented data model a conceptual flow of data is recognized in figure3.

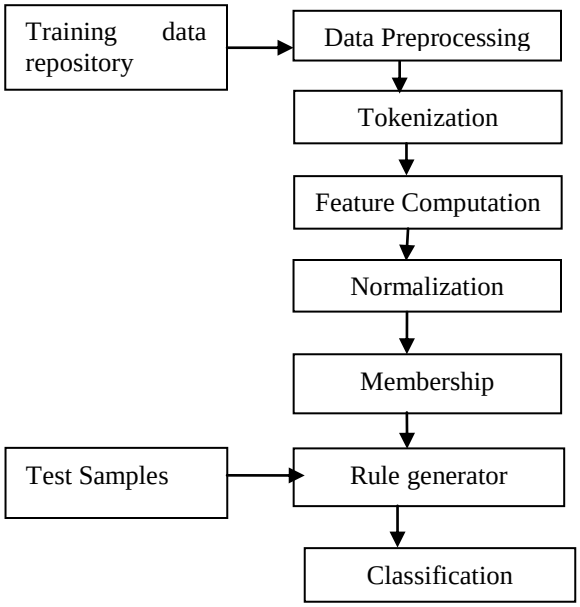

Fig.3: Proposed work on Text Mining using fuzzy logic.

7. CONCLUSION

Text Mining is the technique which is used to extract useful information or knowledge from the text documents which are in the unstructured form. In this paper text mining is discussed with its various techniques used such as summarization, Classification, Clustering, Information Extraction etc. We also discussed about the applications and challenges of text mining. In this paper I also proposed the model which is based on fuzzy logic. The proposed modelincorporate fuzzy logic implementation for distinguish the contents among two given input files and also should becapable to finding similarity among the contents available. Proposed model also used to increaseaccuracy and time consumption during clustering text data.

REFERENCES

[1] Vishal Gupta, S Lehal. " A survey of Text Mining Techniques and Applications". Journal of EmergingTechnologies in web inteliignce, No.1, August 2009.
[2] Vidya k A,G Aghila, "Text Mining Process,Techniques and Tools: an overview",International journal of information technology and knowledge management ,july-december 2010,volume 2,no2,pp.613-622.
[3] Ah-hwee Tan, "Text Mining:The state of the art and the challenges",In proceedings of the PAKDD workshop on Knowledge discovery from advanced databases,pp.65-70,1999.
[4] Vishwadeepak singh baghela, Dr.s.p.tripathi," International journal of computer science issues",vol.9,issue3,pp.545-552,may2012.
[5] Vallikannu ramanathan, T.Meyyappan,"International conference on technology and business management" pp.508-514,March 2013.
[6] Lokesh kumar,Parul kalra Bhatia,"Text Mining :concepts,process and applications " , Journal of global research in computer science ,pp.36-39,march 2013.
[7] Falguni N.patel,Naeha R.soni,"International journal of Advanced computer research", vol.2, pp.243-248, December 2013.
[8] F. Sebastiani, "Machine learning," ACM Computing Surveys, vol. 1, no. 34, pp. 1–47, 2002.
[9] R. Grishman, "Information extraction: Techniques and challenges," in Proceedings of the SCIE, 1997, pp. 207– 220.
[10] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, 2000.
[11] H. Liu, S. B. Johnson, and C. Friedman, "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS," Journal of the American Medical Informatics Associations (JAMIA), vol. 9, pp. 621–636, 2002.
[12] H. Liu, Z. Hu, M. Torii, C. Wu, and C. Friedman, "Quantitative assessment of dictionary-based protein named entity tagging," Journal of the American Medical Informatics Associations (JAMIA), vol. 13, pp. 497–507, 2006.
[13] H. M. Alfawareh and S. Jusoh, "Resolving ambiguous entity through context knowledge and fuzzy approach," International Journal on Computer Science and Engineering (IJCSE), vol. 3, no. 1, pp. 410–422, 2011.
[14] C. Nedellec and A. Nazarenko, "Ontologies and information extraction: A necessary symbiosis," in Ontology Learning from Text: Methods, Evaluation and Applications, P. Buitelaar, P. Comiano, and B. Magnin, Eds. IOS Press Publication, 2005.