ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

An Overview of Text Mining and Its Techniques

[1]M.Kalpana, [2] Dr.S.Nagaprasad, [3] Dr.Manju Khari
[1] Research Scholar-JJT University, Rajasthan,
[2] Faculty of Computer Science, Dept. Of Computer Science, S.R.R.Govt.Arts & Science College, Karimnagar, Telangana State,
[3] Computer Science Department, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi.

Abstract: **- In KDD data Mining step is generally involved with the pattern detection in numeric knowledge, however fairly often necessary (e.g., important to business) data is hold on within the kind of text. Unlike numeric knowledge, text is usually unstructured, and troublesome to handle. Text mining normally incorporates of the analysis of text documents with the extraction of key phrases, concepts, etc. and also the training of the text practice therein way for additional analysis with lot of data Mining approaches. In explorative knowledge Analysis, some initial information is understood regarding the info, however data mining might facilitate in a very a lot of in-depth information regarding the info. Seeking information from huge knowledge is one amongst the primary attributes of knowledge mining. Manual data investigation about for an only some periods currently, however it creates a bottleneck for big data analysis. Variety of information retrieval techniques measure developed to extract this huge quantity of knowledge. Previous studies on data processing specialized in structured knowledge, like relative and transactional knowledge. Conversely, in certainty, a substantial section of the data is hold on in text databases that consist of enormous collections of documents from varied sources, like news articles, books, digital libraries and web content.**

Keywords**: Text Mining, IE, IR, NLP.**

I. INTRODUCTION

Text mining is the concept of discovering knowledge from text and in this a document is usually used as the essential entity of research. A document is probably a series of words and punctuation that is followed by the grammatical guidelines of the language, containing any relevant or applicable phase of text and it may be of any length. The document can either be a paper, an essay, book, web page, emails, etc. on the basis of the sort research that is being performed and relying upon the motives of the scientist. In few cases, a document can probably contain a chapter, one paragraph, or maybe even only one sentence. The most basic unit of text can be defined a word. A term is typically a word; conversely it can still be a word-pair or phrase. Words square measure comprised of characters, and square measure the fundamental units from that that means is built. By using the combination of a word and grammatical structure, a sentence is created. Sentences square measure the fundamental unit consisting of action inside the text and also the info concerning about the action of some subject. Paragraphs square measure the elemental unit of composition and incorporate a connected collection of ideas or actions. Due to the increase in the text length, some of the extra structural forms tend to become relevant that are usually used together with sections, chapters, entire documents, and at last, a collection of documents. And, a lexicon analysis contains entire words that are unique placed inside the corpus [2]. In studies belonging to Text Mining, a sentence is merely treated as a collection of words and also the order in which the words appear will be modified while not affecting the result of the analysis. The grammar structure of a paragraph or sentence is purposely neglected so as to expeditiously handle the text. The bag-of-words construct is additionally stated as interchangeability within the generative language model [1].

1.1. Text Mining Techniques:

Text Mining is a multi disciplinary field that exploits skills that exists in universal field of Data Mining. Moreover, it helps to unite methodologies from a variety of new fields such as Information Retrieval (IR), Information Extraction (IE), Categorization, Topic Tracking, Clustering, Computational Linguistics (CL), Summarization and Concept Linkage [3], [4], [2]. The following sections discuss about these technologies and also the role of each of them in Text Mining.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

1.2. Information Extraction (IE):

According to Data mining research many authors are assumed data or information mined is in the form of structured and in the form of relational databases but many of these information mined in the form of semi structured or unstructured. Information extraction (IE) [6] is a method and it undoubtedly used for retrieval of structured information from the documents that contain only unstructured or semi-structured information. Also, IE make use of NLP to process human language texts. With the help of some predefined sequences that exists in text, also known as pattern matching is used to obtain the final output of the extraction process [5].

1.3 Information Extraction Various Tasks:

Term analysis: This IE task will help to identify frequent word which will occur repeatedly in a document.

Named-entity recognition: This IE task will help to recognize popular terms or keywords which will contain in a document, and mostly it recognizes names of people, organizations and expressions of time.

Fact extraction: This task helps to extract useful relationships from a document.

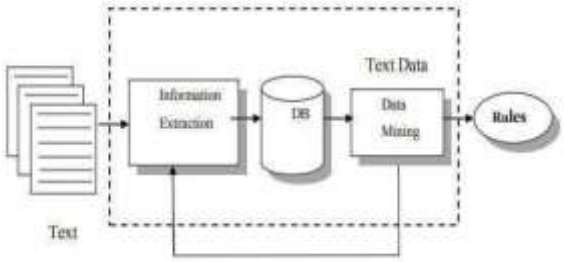

Fig1: Overview of Text mining Framework

1.4. Information Retrieval (IR):

Information Retrieval (IR) is basically the process to retrieve information. With the emergence of text based search engines over the internet, the Information Retrieval (IR) has become a subject of considerable interest. Text is generally considered to be formulated with the help of two fundamental units:

1. A document that can be a book, journal paper, chapters, sections, paragraphs, Web pages, computer source code, and so forth)
2. A term that can be a word, word-pair, and phrase within a document.

In IR, conventionally text queries and documents both are treated as one single entity. This single entity is defined so as to compute the distances between queries and documents which in turn help to provide framework for direct implementation of simple and basic text retrieval algorithm.

1.5. Natural Language Processing (NLP):

NLP is driven form of process approaches for considering and illustrating texts that occur automatically at one or additional levels based on the linguistic analysis in order to achieve the aim of human-like language processing for a various type of applications and tasks. The objective of NLP is to create a computing system for judging, recognizing, and manufacturing human languages. An NLP application comprises the following:

1. AI for modification of human-language text to another
2. To generate human-language text such as fiction, manuals, and documents based on general descriptions
3. To interface alternative systems such as robotic systems therefore sanctioning the employment of human-language kind commands and queries
4. To understand human-language text in order to supply outline or to reach conclusions.

2. Architecture of a Text Mining System:

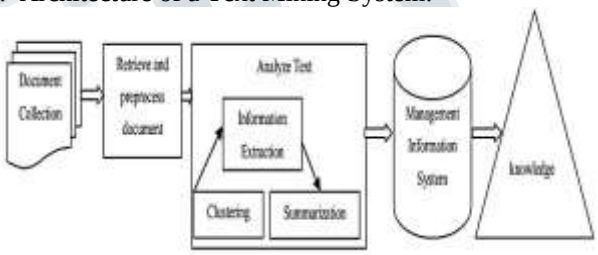

Fig 2: Text mining Architecture

In Text mining process it first takes as set of documents as input and then in second step for preprocessing every document, it checks its format and set of characters [7]. After preprocessing each document, it next move to text analysis phase, in this process may be some steps repeat until the useful information is extracted. And these phases showed in Fig 2. The information thus extracted as a part of the result can be provided as an input for managing information systems and this can yield a significant amount of knowledge of the system for the user. The detailed processing steps are depicted in Fig 3:

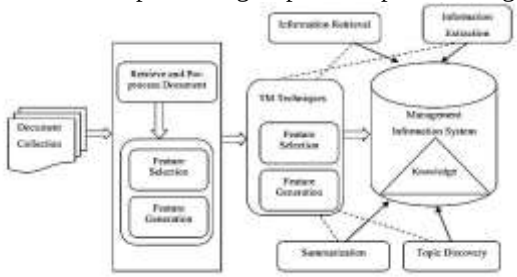

Fig3: Text mining processing steps

Steps followed in Text mining processes that are shown above are briefly discussed below:

- In initial stage various document files of various formats like PDF, txt etc. are obtained from different sources. The document is pre-processed; in this process 3 different tasks are done Tokenize, Feature selection and Feature generation.
- Tokenize task tokenize the document into different entity tokens and it removes noisy, stop words.
- For the representation of unstructured text, generation and selection of features are build based on the retrieve and preprocessed documents. Feature Selection task help to discover the significant features from document.
- After completing above process steps then text mining techniques are applied.

3. Applications of Text Mining:

- Text mining for examining huge collection of unstructured documents, and it can apply for many applications.
- Text mining applications categorized into two ways
- Document analysis Tools: in order to analyze document text content and identifying the relationship between content.
- Document exploration tools: it allows users to search or navigate in document space, and it allows creating clustering in document for similar content.

IV. CONCLUSION

In web 2.0 era a large collections of unstructured documents are generating continuously and most of the conventional techniques are suitable only for structured documents. But data mining task or techniques are not recommended for unstructured so in order to process efficient text mining techniques should applied, in this paper studied different phases of text mining and architecture of text mining, and also applications of text mining.

REFERENCES

[1]. Lee S et.al 2010, "An Empirical Comparison of Four Text Mining Methods", volume 51, Issue. 1, pp. 1-10.

[2]. Stavrianou A et.al 2007, "Overview and Semantic Issues of Text Mining", volume. 36, issue3, pp. 23-34.

[3]. Fan W et.al, 2006, "Tapping into the power of text mining", volume. 49, issue. 9, pp. 76-82.

[4]. Feinerer I et.al, 2008, "Text Mining Infrastructure in R", volume. 25, issue 5, pp. 1-54.

[5]. Gupta V et.al 2009, "A Survey of Text Mining Techniques and Applications" volume. 1, issue 1, pp. 60-76.

[6]. Information Extraction 2011, http://en.wikipedia.org/wiki/Information_extraction.

[7]. Vidhya K et.al 2010, "Text Mining Process, Techniques and Tools: an Overview", volume 2, Issue 2, pp. 613-622.