www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Vandana Publications
IJEMR

Volume-7, Issue-2, March-April 2017
International Journal of Engineering and Management Research
Page Number: 294-296

A Survey on Techniques and Tools of Text Mining

Supreetha Pai[1], Arathi P[2]
[1]Asst. Professor Information Science Department DSATM, Bangalore, INDIA
[2]Asst. Professor Computer Science Department Dr.AIT, Bangalore, INDIA

ABSTRACT

Text Mining is the discovery of new, unknown information, by using the computer automatically for extracting information from different written resources. In this paper, Survey of Text Mining techniques and various tools are presented. Text mining allows you to go further in understanding elements and ideas within a text even better than the search engine, such as Google. Search engines direct the system to search the texts for keywords, but they do not understand the meaning of those words or their context. Text mining techniques enable the user to identify patterns and relationships which exist within a body of texts which would otherwise be extremely difficult or time-consuming to discover. In this paper different text mining techniques for the user data analysis is discussed. Extraction of patterns and arranging the text document is a key goal of text mining technique development. Text mining is related to data mining, except that data mining tools are considered to handle structured data, but text mining can work with formless or semi structured data sets.

Keywords- Text mining, Data Mining, Text Analysis

I. INTRODUCTION

In recent years we are being dependent on the digital / electronic form of data.. For example, when any person buys a ticket or performs any transaction online, his details are stored in the database. Text mining [1] process starts with a document collection from various resources. Text mining tool would retrieve a particular document and pre- process it by checking format and character sets. Then document would go through a text analysis phase. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. Today approx 80% of electronic data is in the form of text This huge data is not only unclassified and unstructured (or semi-structured) but also contain useful data, useless data, scientific data and business specific data, etc.

According to Research 33% of companies are working with high volume of data (i.e. approx 500TB or more). Data mining and Text mining are similar, except data mining works on structured data while text mining works on semi-structured and unstructured data. Data mining is responsible for extraction of implicit, unknown and potential data and text mining is responsible for data.

Text mining process starts with a document collection from various resources. The tools of Text mining retrieves the particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available. They are analysis repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of the System. Text mining process is as shown in following Fig.1.

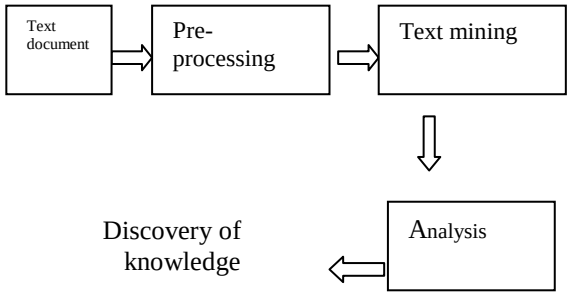

Fig. 1Text mining Process

II. LITERATURE SURVEY

Text analytics first emerged in the late 1990s as "text data mining". Early approaches would treat a text source as a "bag of words." Researchers have adapted statistical techniques to deal with these issues, such as singular-value decompositions and support vector machines for dimensionality reduction coupled with machine-learning algorithms (neural networks and the like) and deeper linguistics that support functions such as semantic disambiguation.

Recently, a new, different style of text analysis has grown in prominence: the use of feature-extraction capabilities have moved beyond entities such as names and e-mail addresses to events and sentiments.

III. MINING TECHNIQUES

According to the information retrieval basically there are four methods used in mining the text data
a) Term Based Method (TBM).
b) Phrase Based Method (PBM).
c) Concept Based Method (CBM).
d) Pattern Taxonomy Method (PTM).

a) Term Based Method

In term based method document is analyzed on the basis of term. These techniques are emerged over the last couple of decades from the information retrieval and machine learning communities. Term based methods suffer from the problems of polysemy and synonymy [6]. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods to solve this challenge.

b) Phrase Based Method

In phrase based method document is analyzed on phrase basis as phrases are less ambiguous and more discriminative than individual terms [7]. The likely reasons for the daunting performance include:
1) Phrases have inferior statistical properties to terms
2) They have low frequency of occurrence
3) Large numbers of redundant and noisy phrases are present among them.

c) Concept Based Method

In concept based terms are analyzed on sentence and document level. The statistical analysis of the term frequency captures the importance of word without document. Two terms can have same frequency in same document, but the meaning is that one term contributes more appropriately than the meaning contributed by the other term [8].
Concept-based model can effectively discriminate between non important terms and meaningful terms The concept-based model usually relies upon natural language processing techniques.

d) Pattern Taxonomy Method

In pattern taxonomy method documents are analyzed on pattern basis. Patterns can be structured into taxonomy by using is-a relation. Patterns can be discovered by data mining techniques like association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining [9]. Use of discovered knowledge (patterns) in the field of text mining is difficult and ineffective, because some useful long patterns with high specificity lack in support and yield a ineffective performance.

IV. TOOLS FOR TEXT MINING

The various text mining tools [2] are described here such as Wordstat, QDA miner, open NLP, Gensim,Carrot-2,Languageware, ABBYY, Angoss, Attensity, NLTK etc.

a) **WordStat** is software that helps people analyze large amounts of written documents like customer surveys, political speeches,, emails or twitter chats. The software helps people find common topics, themes and hidden meanings in unstructured text data. The software helps search those responses and generate common themes or topics to target where they need to improve. For example hotel chain may have 100,000 customer comments about their rooms, food service or outdoor facilities.

b)**QDA Miner** is a qualitative tool, providing powerful coding, annotating, retrieving and analyzing functions. They provide sophisticated and flexible means to integrate qualitative and quantitative data and analyses, in ways unique in comparison to other options currently available.

c) The **Apache Open NLP** library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co reference resolution. These tasks are usually required to build more advanced text processing services.

d) **Gensim** is a open-source Python library which is designed to handle large text collections, using data streaming and efficient incremental algorithms. They analyse Semantic structure of the document.

e) Carrot-2 is an open source tool that can automatically cluster small collections of documents, e.g. search results or document abstracts, into thematic categories.

f) Language Ware is a natural language processing (NLP) technology developed by IBM, which allows applications to process natural language text.

It comprises of Java libraries which provides the language identification, normalization, entity and relationship extraction, and semantic analysis and disambiguation, , text segmentation/tokenization

g) ABBYY Compreno is a natural language processing technology that applies morphology, syntactic and semantic text analysis to extract insights and intelligence from unstructured content.

h) Angoss provides entity and theme extraction, topic categorization, sentiment analysis and document

summarization capabilities via the embedded Lexalytics Salience Engine.

i) Attensity is hosted, integrated and stand-alone text mining (analytics) software that uses natural language processing technology to address collective intelligence in social media and forums; the voice of the customer in surveys and emails; customer relationship management; e-services; research and e-discovery; risk and compliance; and intelligence analysis.

j) The **Natural Language Toolkit**, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language.

V. APPLICATIONS IN TEXT MINING

Text mining applications are being used in all the fields such as health care, politics, banks, IT sector, research, energy, media, political analysis etc, to get most valuable and useful data from an enormous amount data.

Some of the applications of text mining are mentioned below:

a) Text mining can be used in collecting all possible information about market trends and other competitors to yield competitive advantage [10].

b) It can be used to detect unwanted junk e-mails automatically.

c) It can be used to manage human resources. It can be used to analyze the staff's opinion, monitoring company's progress and monitoring satisfaction levels of employees.

d) By using text Analysis Company can provide customer relationship management to provide quick response to any client's query or message.

e) Text mining techniques can be used to analyze different web pages in a different language [10].

f) Normally people like to see headlines in a newspaper which involves naming of any person, country or organization. Text mining techniques can be used for the classification of news as Text.

g) Text mining can also be used for Sentiment Analysis. It is a case of natural language processing which could mark the mood of the people about any specific product by analysis and classifying it as positive, negative or neutral. Sentiment Analysis is a process of automatic extraction of features by mode of notions of others about specific product, services or experience.

VI. CONCLUSION

Text mining applications are vide and plays vital role in various sectors like publishing and media, telecommunications, Internet, Banks, insurance and financial markets, public administration, healthcare. Text mining is one of the fastest growing fields today. With the passage of time its importance is only going to increase because rate of data production is very high. The growing interaction of text mining to other fields, especially with machine learning, visualization and natural language processing, has made it is possible to design more effective and useful text mining system. In this paper we tried to present an overview of text mining approach with its techniques, tools and applications.

REFERENCES

[1] Kaushik A.,Naithani K. ,A Comprehensive Study of Text Mining Approach .IJCSNS International Journal of Computer Science and Network Security (2016).

[2] Jan van Gemert, "Text Mining Tools on the Internet: An Overview".

[3] Simona Balbi, Sergio Bolasco and Rosanna Verde, "Text Mining on Elementary Forms in Complex Lexical Structures"

[4] http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis.

[5] http://www.cio.com.au/article/575209/5-tools-techniques-text-analytics/

[6] G. Salton and C. Buckley, "Term-Weight Approaches in Automatic Text Retrieval," Information Processing and Management: An Int'l J., vol. 24, no. 5, pp. 513- 523, 1988.

[7] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,"

[8] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[9] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[10] Divya Nasa, "Text Mining Techniques- A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.

[11]E-book: Introduction to Data Mining and Knowledge Discovery Third Edition By Two Crows Corporation, ISBN: 1892095-02-5.