International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241 Volume: 4; Issue: 2; February -2015

A Survey on Text Mining Approaches

Saroja Devi H [*1] *,Mahesh Kini M* [#2]
[*1] *Professor and HOD, Computer Science and Engineering, NMAMIT, Nitte.*
[#2] *PG Student, Computer Science and Engineering, NMAMIT, Nitte.*
[1] hsarojadevi@gmail.com
[2] *mahesh_nitte@yahoo.com*

***Abstract*—Text mining is a technique to discover or find interesting patterns from the available text documents, which is more prevalent in text mining applications. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text. The pattern discovery from the collection of text documents is a well-known problem in text mining. Analysis of terms or text content and categorization of the documents is a complex task of data mining. Various techniques of text categorization and classification have been developed in order to efficiently and effectively handle the task of text mining. Some of them are based on supervised and others unsupervised methods. Various approaches to carry out text mining are presented in this paper.**

Keywords - Text mining, Classification, Categorization, Survey.

I. INTRODUCTION

Rapid growth of text mining as a technology for analyzing large volumes of unstructured textual documents, with the "information explosion" is an accepted fact in the 21st century. This brings in thrust towards addressing the problem of text mining related to critical data, which is overlooked. Further, due to computational automation in many fields, various different sources of text documents in different formats are extensively used and made available for information exchange. Hence there is a greater need to categorize, classify and deal with other technologies with the document for further processing. Extraction of patterns that are interesting from such documents and arranging the text document are among main goals of text mining technique that is to be designed and developed. Text mining is related to data mining, except that data mining tools normally preferred to handle structured data, whereas text mining deals with unstructured or semi-structured data sets. The application of text mining is very popular in email analysis, opinion mining, digital libraries, and medical report analysis and so on.

The text mining techniques begin with collection of text documents (text repository), where a text mining tool for pre-processing is applied. Pre-processing technique mainly deals with cleaning and formatting the data, in addition to extracting the meaningful features from the documents. In the next step the text mining techniques such as classification or clustering algorithm [16] can be applied to arrange the documents. The text mining methods have become emerging technologies in the area of Web Intelligence where sources of document are websites or web servers.

In our paper, we have discussed various approaches of text mining. While references [1], [4], [10], [14] and [17] also do a survey of the related concepts and our paper included the recent approaches for text mining as well. Further, classification tree and comparative study of approaches is done which is a key contribution to growing field of text mining. In addition, we briefly discuss text mining approaches to the number of recent research and application oriented areas used presently and in future. Text mining comparison tools based on different approaches are given to direct the readers.

The basic text mining technologies presented in survey papers in the recent past are studied and highlighted with comparison. In this paper first section of the paper presents the general overview of text mining and their applications. The next section briefs the recent research and development over the text mining techniques. In section III our paper highlights the mainly used text mining technologies that are presented in the different papers. Section IV explores the latest and recent trends in text mining approaches.

II. Text mining

Text mining is the process of computation that involves extraction of information from bulk quantity of data and uncovering new unidentified information by retrieving from numerous written and digital resources with the help of algorithms of robust form. Various approaches of text mining are surveyed and presented.

The steps involved in the text mining process flow are depicted in the Fig 1 [1].

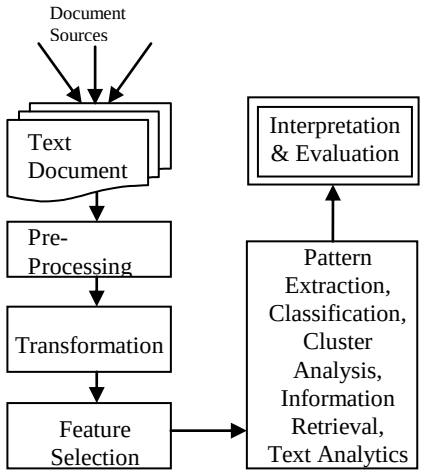

Fig.1. Text mining process flow.

A. Text Pre-processing

The text pre-processing step is further divided into number of sub tasks as follows:

i) Tokenization: Text document, collection of sentences divided into words by removing spaces, commas or any other delimiters.

ii) Stop word Removal: This step involves removal of common words like 'a', 'of' or any other tags in the collection of tokens.

iii) Stemming: This technique is used to find the root or stem of a word. Stemming converts words to their root words. For example the words like ran, running converted to run [2].

B. Text Transformation or Feature Generation

Converting text document into vector space termed as text transformation which can be used for further analysis task effectively.

C. Feature Selection/Attribute Selection

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required.

D. Text mining methods

Number of text mining methods in data mining had been proposed such as: Classification, Clustering, Information retrieval, Topic discovery, Summarization, Topic extraction.

E. Interpretation or Evaluation

This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

III. TEXT MINING TECHNOLOGIES

Vishal Gupta et. al. in [4] has proposed efficient technologies which explore that Language analysis, understanding, Generation of text etc. can well implement in computer, which work as human being. Recent Technologies can achieve this using following methods. In this section we discuss all technologies with example so it will be useful to work further in the interested area.

A. Information extraction (IE)

Paper [5] discusses Information extraction technique which finds key phrases and relationship within a text. For that it uses pattern matching method. Pattern matching involves matching predefined sequences of text with text considered for test. This technique is very useful in analyzing large text dataset. The extracted information by IE cannot be represented directly into a structured form. Hence post processing is required. The paper presents a framework for text mining, called DISCOTEX (discovery from text extraction), using a learned information extraction system to transform text into more structured data which is then mined for interesting relationships.

B. Classification

Paper [6] discusses a very innovative method of classification technique that classifies text documents into predefined class label (categories). Classification has been used in many applications like Email or Mobile message classification, online customer feedback classification, business reports classification etc. Classification can be integrated with topic tracking to classify the documents by topic and thus making the process faster. This paper describes Naïve Bayesian (NB) approach for the automatic classification of websites based on content of home pages. The NB approach, is one of the most effective and straightforward method for text document classification and has exhibited good results in previous studies conducted for data mining.

Classification of web content is different in some aspects as compared with text classification. The uncontrolled nature of web content presents additional challenges to web Page classification as compared to traditional text classification. The web content is semi structured and contains formatting information in form of HTML tags. A web page consists of hyperlinks to point to other pages. This interconnected nature of web pages provides features that can be of greater help in classification.

C. Summarization

The paper [7] presents efficient summarization technology that condenses the source text into a shorter version preserving its implied meaning of information. Human cannot manually summarize large documents [5]. Highlighting summary with main points reduces the time required to spend in reading all documents by the researcher in research organizations or institutes. Text Summarization methods can be broadly classified into extractive and abstractive summarization. An extractive summarization method involve selecting important words, sentences, paragraphs etc. from the original document and reduce them into shorter form. The importance of sentences is decided based on statistical [6] and linguistic features of sentences. An abstractive summarization attempts to develop a meaning out of main concepts in a document and then expresses those concepts in natural language. It uses linguistic methods to examine and interpret the text and then to find the similar concepts and expressions to best describe it by generating a new shorter text that conveys the main important information from the original text document. One of the strategies most widely used by text summarization tools is statistically weighting the sentences. Other strategy may be rule reduction method which makes use of grammar rules. Then features of alphabet tokens are identified like Determiner, Preposition, Noun, Verb, Adjective etc. This is done based on the rules we have defined in the text analyser. Finally using the rules, the analyzer categorizes the tokens into Noun Phrase, Possessive Pass, Prepositional Phrase or Verb Phrase based on its feature (noun, verb, preposition etc.) and then summarizes them to formulate a sentence.

D. Topic Tracking

As proposed in [8], an innovative text mining technique, *topic tracking* deals with maintaining the topics that are already referred by the user to facilitate user's new search

with prediction of topics automatically with the previously searched topics. Topic detection studies the problem of detecting new and upcoming topics in time ordered documents related to multilingual news oriented textual topics. The methods are frequently used in order to detect and monitor news tickers or news broadcasts. There are number of areas where topic tracking can be useful in industries. It can be used to alert production companies about their competitors in the news. This allows them to keep up with competitive products or changes in the market.

In this technique keywords are extracted as a set of significant words in an article that gives high-level description of its contents to readers. Manual keyword extraction is an extremely difficult task and time consuming. For a rapid use of keywords, we need to implement an automated process that extracts keywords from news articles. The process of keyword extraction model is shown in fig 2.

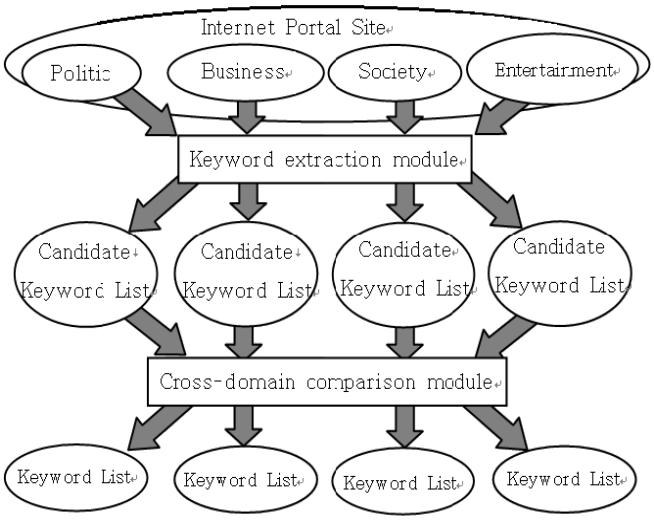

Fig 2. The architecture of keyword extraction system

HTML news documents are collected from the Internet portal sites. Candidate keywords are extracted by keyword extraction module. Finally keywords are extracted by cross-domain comparison module. Keyword extraction module creates a set of tables for 'document', 'dictionary', 'term occur fact' and 'TF-IDF weight' in relational database. In the beginning, collected news documents are stored in 'Document' table and nouns are extracted from the documents in 'Document' table. Next, TF-IDF weights for each word are calculated using 'Term occur fact' table and the result are updated to 'TF-IDF weight' table. Finally, using 'TF-IDF weight' table, 'Candidate keyword list' for each news domain is prepared with words is ranked high. Multi vector or Lexical chaining method can be applied for topic tracking which involves tracking a given news event in a stream of news stories i.e. finding all the related stories in the news stream.

E. Clustering

As discussed in [9], clustering is a technique in which there are no predefined class labels. While classifying text documents clustering technique uses similarity measures between different entities and the most similar entity belong to one class and dissimilar belong to another class. A basic clustering algorithm creates a vector of topics for each document and measures the term weights [15] of how best the document fits into each cluster. Clustering technology is useful in the organization managing information systems which may contain thousands of documents. In K-means clustering algorithm, similarity between text documents is calculated, not only based on eigenvector frequency statistics, but also combine the degree of association between words , and then the relationship between keywords has been taken into consideration.

In fig 3 the general steps used in document clustering are described. Initially words are separated and then weights are applied to each of them. Then similarity measures are determined and finally different effective clustering algorithms like k-Means or k-Medoids Partitioning can be applied.

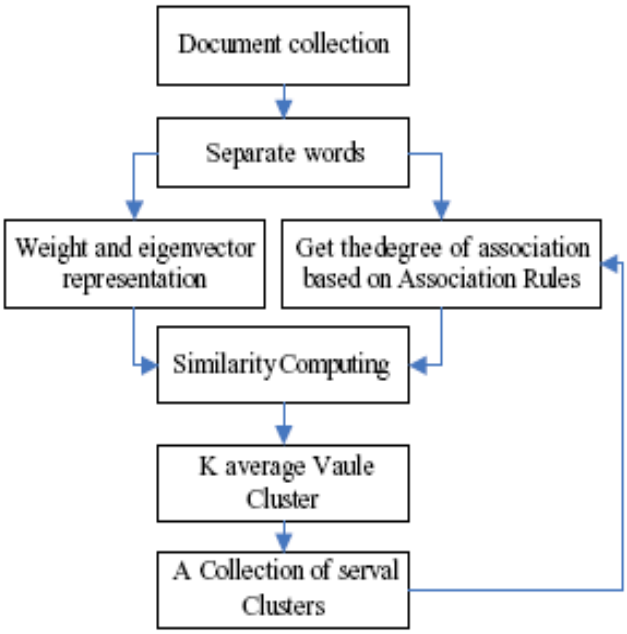

Fig 3. The general flow of k-means clustering

F. Concept Linkage

According to [10], the idea of Concept linkage connects related documents by identifying shared concepts between two unrelated data sets. The unexpected connections may become apparent by examining these shared concepts. The primary goal of concept linkage is to provide connections for information rather than searching for it as in information retrieval. For example in biomedical report analysis, concept link used to link diseases and treatment. Concept linkage tools link related documents by identifying their related or shared concepts and help users find meaningful information that they have not found using traditional searching methods.

It indicates browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the field of biomedical where research has been done that it is impossible for researchers to read all the terms and make associations to other related documents. Preferably, concept linking tool can identify links between diseases and treatments when cannot manually. For example, a text mining software solution must easily imply a link between topics X and Z, when there is a well-known link between X and Y & Y and Z.

G.Natural Language Processing (NLP) or Computational Linguistics

As discussed in [11], natural language processing (NLP), is the attempt to extract a more complete 'meaning representation' from free text which often involve text mining. The goal of NLP is to design and build models which involve a clear classification of various linguistic concepts (e.g. full parsing vs. shallow parsing vs. Heuristics to get information or concept) that will analyze, understand and generate NLP. This can be put roughly as figuring out who did what to whom, when, where, how and why. Also application includes machine translation of one natural language text to another. Thus it is useful for enabling the use of natural language for providing a summary after understanding any text document, commands or queries for further analysis purpose, thus to develop a linguistic analytic model. NLP typically makes use of linguistic concepts such as part-of-speech (noun, verb, adjective, etc.) and grammatical structure (either represented as phrases like noun phrase or prepositional phrase, or dependency relations like subject-of or object-of). It has to deal with anaphora (what previous noun does a pronoun or other back-referring phrase correspond to) and ambiguities (both of words and of grammatical structure, such as what is being modified by a given word or prepositional phrase). To do this, it makes use of various knowledge representations, such as a lexicon of words and their meanings and grammatical properties and a set of grammar rules and often other resources such as ontology of entities and actions, or a thesaurus of synonyms or abbreviations.

H.Text-mining approaches in Molecular Biology and Biomedicine

As in paper [17], Martin Krallinger et. al., the field of NLP is concerned with the analysis of free textual information and applied in the context of molecular biology. Text-mining approaches involve extracting and analyzing information from large collections of free textual data by using automatic or semiautomatic systems. Currently, text-mining applications are being employed in the identification of biological entities such as protein or gene names, automated protein annotation, analysis of microarrays and extraction of protein–protein interactions. In general, text-mining applications take advantage of a range of domain-independent methods such as part-of-speech (POS) taggers, which label each word with its corresponding part of speech (e.g. noun, verb or adjective), or stemmers, which are algorithms that return the morphological root of a word form. Also, domain-specific tools and resources such as protein taggers and ontology are employed.

I. Question Answering System

In [18], a simplified and innovative technique of the Question Answering model for restricted domain uses advanced NLP tools which has text mining techniques and basically works over the concept of Information Extraction rather than the old technique of information Retrieval used by the search engines. The main objective of the model is to extract the exact and precise answer for the given question from a large dataset. This framework is simple and easy to implement. The Framework is divided into four modules namely: Question Processing, Document Processing, Paragraph extraction and Answer extraction modules. The model consists of algorithms separately for Definition, Descriptive and Factoid types of questions for extracting most potential answer from the large dataset. Question Answering model can be define as: "A system, whose main objective is to determine WHO did WHAT to WHOM, WHERE, WHEN, HOW and WHY?".That is a system which is capable of answering all the WHO, WHAT, WHOM, WHERE, WHEN, HOW and WHY type questions of the user. The process involves the use of semantic information specifically in the answer extraction step [15]. A model scheme is shown in fig.4.

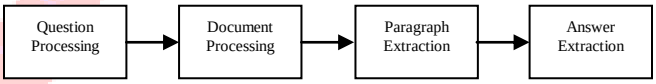

Fig4. Question Answer flow model.

J. Classification tree structure:

Text classification tree shown below clearly distinguishes different approaches for text classification. We also observe that the different algorithmic techniques for each classification types. Statistical based classifications make use of probability theory as base and rule based works on natural language processing. Hybrid approach combines several of the statistical and rule based to make the classifier predict better and more efficient. Fig.4 represents pictorial observations as sho

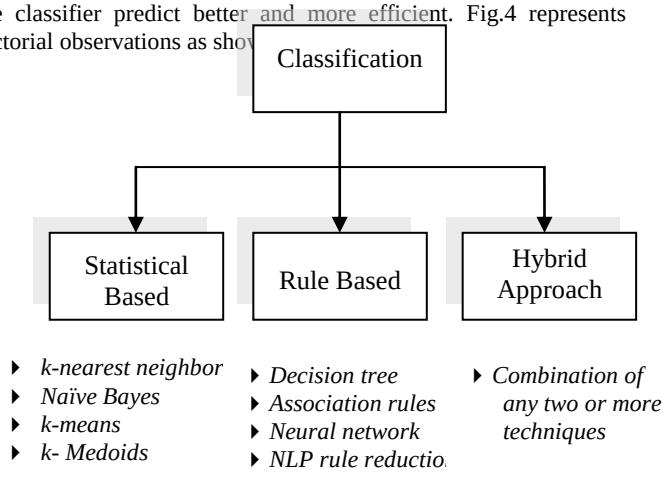

Fig 4.Classification tree structure

K. Text Mining Techniques Comparisons:

Text mining uses different approaches which play an important role. The algorithms differ from each other. The unstructured text information retrieved are basically improve the knowledge base so that the prediction process improves over the period of time from structured database. The Summarization technique is used to condense the document which reduces length of the document without changing the most important meaning. The categorization is supervised course of action and uses pre-labeled set of documents according to their contents. While as the clustering is used to find fundamental structures in information, and rearrange them into related groups for further analysis. It is an unsupervised process using which entities are classified into groups called clusters, and hence clustering technique. Clustering deals with towering dimensional data, finding interesting patterns associated with data extracted. Another feature is that it is a group of similar type of data and their relationship between them.

Table 1: Comparing Text mining techniques

Technique	Process Behaviour characteristics	Tools/Algorithms/ Techniques
Classification	Document and Text, based on Supervised or unsupervised methods.	Naïve Bayes Decision trees
Clustering	Grouping, Collection of documents, Analysis and Classification of text document.	k-means, k-Medoids, Rapid Miner
Categorization	Word, phrases and sentence based on Document.	Intelligent Miner based on statistical or NLP based
Summarization	Condense by sentence or paragraph, keeping its meaning and overall implications as is.	Tropic Linking/Tracking Tool, Rule reduction tool
Extraction	Retrieve meaning information from unstructured text.	Text Finder, Clear Forest Text
Retrieval	Information of interesting patterns from text documents for further analysis from text document.	Intelligent Miner, Text Analyst

IV. RECENT TRENDS

Efforts are still required in developing systems that interpret natural language queries and automatically performs the appropriate mining operations. Text mining now used in security purpose, which “bug” or “roomer” sms are detected on mobile system, classified and eventually removed. Therefore in this context, sms classification is also required more future work in this area of text mining. A social classification solution (SCS) model has been built by Government and corporate agencies to understand the critical problems that are faced by the public through the social websites, anonymous calls or smses. SCS model can classify text on area based and profession based to understand the common problems effectively. Use of text mining is a major challenge in these fields. Web crawler (or classifier) tool has been designed to extract relevant documents from websites and severs.

V. CONCLUSION

Text mining, also known as Text Data Mining or Knowledge-Discovery in Text (KDT) [13], refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Its aim is to get insights into large quantities of text data. We also discuss the General Process of Text encoding and mining. Current text mining products and applications are designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. The growth of web technologies has lead to a greater interest in the context of classification of text documents containing links or other information.

VI. REFERENCES

[1] Falguni N. Patel, Neha R. Soni, “Text mining: A Brief survey”, International Journal of Advanced Computer Research (ISSN) Volume-2 Number-4 Issue-6,243-248, Dec-2012.

[2] M. Porter. “An algorithm for suffix Stripping and stemming”. A Text Book, pages 130–137, 1980.

[3] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[4] Vishal Gupta and Guruprit Lehal, “A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, 60-76, August 2009.

[5] N. Kanya and S. Geetha ,“Information Extraction: A Text Mining Approach”, IET-UK International Conference on Information and Comm. Technology, IEEE, Dr. M.G.R. University, Chennai, India, 1111- 1118, 2007.

[6] Ajay S. Patil, B.V. Pawar, “Automated Classification of Web Sites using Naive Bayesian Algorithm”, Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong Vol I, 14-16, 2012.

[7] C. Lakshmi Devasenal and M. Hemalatha, “Automatic Text Categorization and Summarization using Rule Reduction”, IEEE-

International Conference on Advances In Engineering Science & Management,594-598, March 2012.

[8] Sungjick Lee & Han-joon Kim, "News Keyword Extraction for Topic Tracking", 4th International conference on Networked Computing and Advanced Information Management, IEEE, Korea. 554-559, 2008.

[9] Liritano S. and Ruffolo M., "Managing theKnowledge Contained in Electronic Documents: a Clustering Method for Text Mining", IEEE, 454-458, Italy, 2001.

[10] Mr. Rahul Patel , Mr. Gaurav Sharma "A survey on text mining techniques", International Journal Of Engineering And Computer Science,Volume 3 Issue 5, 5621-5625, May-2014.

[11] U. Ackermann, B. Angelini, F. Brugnara, M. Federico, D.Giuliani, R. Gretter, G. Lazzari and H. Niemann, "Speed Data: Multilingual Spoken Data Entry", International Conference, IEEE, Trento, Italy, 2211-2214.

[12] PayaJ Biswas, Aditi Sharan, Nidhi Malik,"A framework for restricted domain question answering system", IEEE International Conference (ICICT), New Delhi, 613-620, 2014.

[13] Shaidah Jusoh and Hejab M. Alfawareh "Techniques, Applications and Challenging Issue in Text Mining", IJCSI International Journal of Computer Science Issues, Saudi Arabia Vol. 9, Issue 6, No 2, Nov-2012

[14] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay3, "A tutorial review on Text Mining Algorithms" [International Journal of Advanced Research in Computer and Communication Engineering. Vol. 1, Issue 4, June-2012.

[15] R.C. Balabantaray, D.K. Sahoo, B. Sahoo, M. Swain, "Text Summarization using Term Weights", International Journal of Computer Applications Volume 38– No.1, 10-14, Jan-2012.

[16] Shady Shehata, Member, Fakhri Karray, Senior Member and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, Oct-2010.

[17] Martin Krallinger, Ramon Alonso-Allende Erhardt and Alfonso Valencia, "Text-mining approaches in molecular biology and biomedicine", DDT, Volume 10, Number 6, March-2005.

[18] Moreda P., Llorens H., Saquete E., & Palomar M., "Combining semantic information in question answering systems", Information Processing & Management, 870-885, 2011.