# A Review of Text Mining Techniques Associated with Various Application Areas

**Dr. Shilpa Dang[1], Peerzada Hamid Ahmad[2]**

[1]Assistant Professor, MMICT&BM, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India

[2]Research Scholar, MMICT&BM, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India

**Abstract:** ***Text mining is used to extract hidden valuable information from semi-structured or unstructured. The amount of information available is day by day increasing at a dramatic rate. This reality has led to investigate various text mining techniques. In this paper our focus is to review the basic concept of various text mining techniques and its applications. We present the basic differences between relative terminologies on the basis of motivation, process and model used and the algorithms used. In addition to this we have also discussed a comparison between text mining techniques on the basis of characteristic, algorithms used, models used and tools.***

**Keywords**: Categorization, Clustering, Information Extraction, Information Retrieval, Summarization.

## 1. Introduction

Today the web is the main source for the text (documents), the amount of textual information available to us is consistently increasing. Approximately 80% of the information of an organization is stored in unstructured format (reports, email, views and news etc.) This shows that approximately 90% of the world’s data is held in unstructured formats. The need of automatically retrieval of useful knowledge from the large amount of textual data in order to assist the human analysis is fully apparent [1]. Increasingly, however, large amounts of information such as textual information are unstructured, and defy simple attempts to make sense of it. Manual analysis of this unstructured textual information is increasingly impractical, and as a result, text mining techniques are being developed to mechanize the process of analyzing this information.

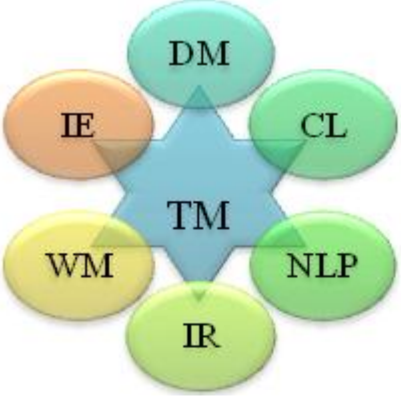

**Figure 1.1:** Text Mining Basic Areas

Text Mining [2] is the finding previously unknown hidden information. The information extracted from different written resources is done automatically. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar within web search. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data items and the quantitative methods used to analyze these data items [3]. The fundamental objective of text mining is to enable users to extract data from text based assets and manages the operations like retrieval, extraction, summarization, categorization (supervised) and clustering (unsupervised). Text mining is the young interdisciplinary field which is incorporated with data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing as shown in Figure 1.1.

The rest of the paper is organized as follows. Section II introduces introduction of text mining techniques and its applications. Section III gives the differences between the relative terminologies. Section IV gives the comparison between text mining techniques. Finally, the conclusions are made in Section V.

## 2. Techniques Used in Text Mining

Text mining is an interdisciplinary field that utilizes techniques and combines methodologies from various other areas such as information retrieval, information extraction, text categorization, text summarization and text clustering. In this section we will discuss each of these technologies and its application they play in the text mining.

### 2.1 Information Retrieval

Information retrieval is a relatively old research area. It gained increased attention with the rise of the World Wide Web and the need for sophisticated search engines. The most well known information retrieval (IR) systems are search engines such as Google which identify those documents on the World Wide Web that are relevant to a set of given words. For instance, Google tries to find a set of available documents on the web using a search phrase. It tries to find matches for the search phrase or parts of it [4]. The pre-processing work for the search engines is the information extraction process to create order in a chaos of information. Google crawls the web for information, interprets it and stores in a specific structure so that it can be quickly accessed when users are firing search phrases. Information retrieval is the task of obtaining relevant information from a collection of various resources [5]. It is used to focus on the textual information which includes text as well as document retrieval (web pages, pdf's, pp slides, paragraphs, articles etc.).

Document retrieval is measured as an extension of the information retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Each user tries to locate documents that can give up information required and tries to satisfy information needs for a particular user. The process of acquiring, identifying and searching the possible documents that may meet this information need is called retrieval process [6]. All of the retrieved documents intend to satisfy user information needs expressed in natural language text. To studies the retrieval of information from a collection of written text documents is called Information retrieval (IR). Therefore information retrieval (IR) can be defined as a set of methods and techniques for formulating information needs of the users in form of queries. The query is then used to select a relevant document from a larger collection database (web). It can reduce information overload by using automated information retrieval systems. The information retrieval mostly deals with the large range of information processing from retrieval of information to the retrieval of knowledge. This system is used by many universities, public libraries, government and companies to provide access to articles, books, journals and other documents [7], [8].

### 2.1.1 Application Area

It is widely used now-a-days in retrieving information (speech, images, videos, news, blogs, music etc.). Image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Music information retrieval is the growing field of research with many real-world applications in retrieving information from music. Another application is in the field of digital repository which is a focused collection of digital objects that can include text, visual material, audio material, video material, stored as electronic media formats along with means for organizing, storing and retrieving the files and media contained in the library collection [1], [7]. It is also used in Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi) automated or computerized methods prior to presentation to a human user. Its main goal is the management of the information overload and increment of the semantic signal-to-noise ratio. Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. It is also used in geographic, chemical structure, software engineering etc. to retrieval information [8], [9].

### 2.2 Information Extraction

Information extraction (IE) is the task of automatically extracting structured specific information from unstructured or semi-structured natural language (in the form of text). It recognizes the extraction of entities such as names of persons, organisation, location and relationship between entities attributes events and relationships from text. The valuable information extracted is without proper understanding of text such as name of a person, organisation, location and sex [5], [10]. These are stored in database like patterns and are then available for further use. In most of the cases this activity concerns processing human language texts by means of processing of natural text language.

The information gathered is well-organized (structured) and stored in a database automatically. IE transforms a corpus of textual information into a more structured database. The database constructed by IE module then can be provided to the KDD module for further mining of knowledge as shown in the Figure 1.2. Its complexity of in use methods depends on the features of source texts. The method can be rather simple and definite if the source is well structured. If the source of information is less ordered or even plain text language (natural), the complexity of the this method becomes high as it includes natural language identification and analogous processes. The major advantage of information extraction systems is the accuracy of the queries and the clearness of the output. They can be efficiently reviewed and then entered into a database or displayed visually on screen [11].

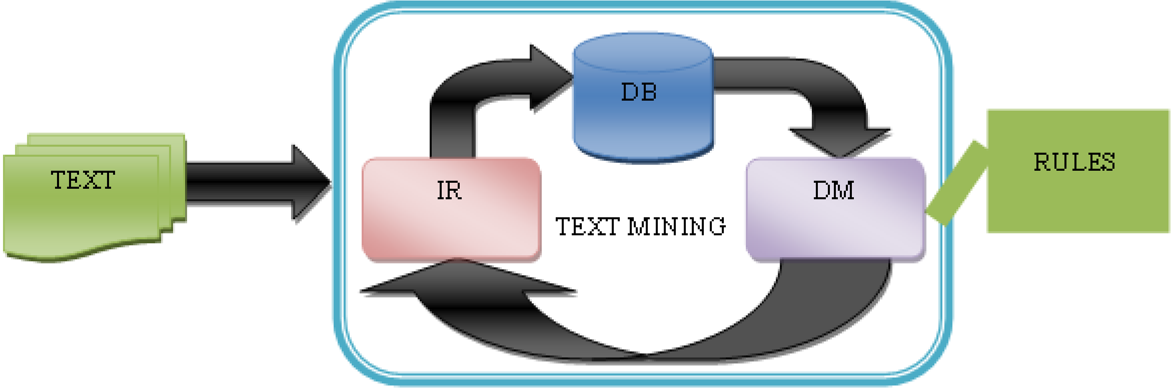

**Figure 1.2:** IE based Text Mining Framework

### 2.2.1 Application Area

It is useful for a variety of applications particularly given the recent proliferation of internet and web document. Recent applications include job posting and resources, medical patient records, weather reports, seminars announcements, course homage and apartment rental ads. In digital libraries metadata is structured data for helping users find and process documents and images. It is used to email data. Email is one of the commonest means for communication via text. Many text mining applications need take emails as input like email analysis, email routing, email filtering, IE from email and newsgroup analysis. It is used to text block detection (header, signature, program code, quotation, paragraph detection) and block-metadata detection (header, signature etc.). It also used to extract information in person information management (person profile, contact information etc.) [2], [13].

### 2.3 Text Summarization

The definition of the summary is an obvious one which emphasizes the fact that summarizing is in general a hard task because we have to characterize the source text as a whole and capture its important content. The content is a matter of both information and its expression and importance is a matter of what is essential as well as what is salient [12]. Summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown and as the quantity of data has increased, so has interest in automatic summarization. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. An example of the use of summarization technology is search engines such as Google and another is the document summarization. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs. Summarization of multimedia documents, e.g. pictures or movies is also possible. Some systems will generate a summary based on a single source document, while others can use multiple source documents. These systems are known as multi-document summarization systems [14], [16].

#### 2.3.1 Application Area

The application areas for text summarization are extensive. As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. Information is published simultaneously on many media channels in different versions like a paper newspaper, web newspaper, WAP newspaper, SMS message, radio newscast and a spoken newspaper for the visually impaired. Customisation of information for different channels and formats is an immense editing job that notably involves shortening of original texts. It is also used to produce a draft summary. Documents can be made accessible in other languages by first summarizing them before translation, which in many cases would be sufficient to establish the relevance of a foreign language document. Automatic text summarization can also be used to summarize a text before an automatic speech synthesizer reads it, thus reducing the time needed to absorb the key facts in a document. In particular, text summarization can be used to prepare information for use in small mobile devices, such as a PDA, which may need considerable reduction of content. Image collection summarization is other application example of automatic summarization [13], [15], [18].

### 2.4 Text Categorization

Categorization is the process of assigning a given text into groups of entities whose members are in some way similar to each other. Recognition of resemblance across entities and the subsequent aggregation of like entities into categories lead the individual to discover order in a complex environment. Without the ability to group entities based on perceived similarities, the individual's experience of any one entity would be totally unique and could not be extended to subsequent encounters with similar entities in the environment. This process is considered as a supervised classification technique since a set of pre-classified documents is provided as a training set. The goal of TC is to assign a category to a new document. By reducing the load on memory, facilitating the efficient storage and retrieval of information, categorization serves as the fundamental cognitive mechanism that simplifies the individual's experience of the environment [2], [17].

#### 2.4.1 Application Area

TC can play an important role in a wide variety of areas such as information retrieval, word sense disambiguation, topic detection and tracking, web pages classification, as well as any application requiring document organization. Automatic indexing of articles (scientific) by way of a controlled vocabulary such as the classification scheme (ACM) where the categories are the entries of the controlled vocabulary. Recently text categorization has aroused a lot of attention for its possible application to automatically classifying sites or Web pages under the hierarchical catalogues hosted by popular Internet portals. Classifying Web pages automatically has understandable advantages [7], [6].

Another application of text categorization is the word sense disambiguation which is the activity of finding given the occurrence in a text of an ambiguous word and the sense of this particular word occurrence. Filtering system (document, e-mail, filters of newsfeed or unsuitable content filters) is the bustle of classifying a flow of incoming information dispatched in an asynchronous way by a producer (information) to a consumer (information) [18]. A document which is not likely interested by the consumer should be blocked by the filtering system. The explosion in the availability of digital information has boosted the importance of such systems, which are nowadays being used in diverse contexts such as the creation of personalized Web newspapers, junk e-mail blocking and Usenet news selection.

### 2.5 Text Clustering

Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". The idea of clustering originates from statistics where it was applied to numerical data. However, computer science and data mining in particular, extended the notion to other types of data such as text or multimedia [1], [19].

Clustering is an unsupervised process through which objects are classified into groups called clusters. In the case of clustering, the problem is to group the given unlabeled collection into meaningful clusters without any prior information. Any labels associated with objects are obtained solely from the data. An advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results.

### 2.5.1 Application Area

Clustering is used in a wide variety of scientific fields, applications and data analysis fields including data mining, document retrieval, image segmentation and pattern classification. The application of document clustering can be categorized to two types online and offline [19]. Online applications are usually constrained by efficiency problems when compared offline applications. Biologist has applied clustering to analyse large amounts of genetic information and find groups of genes that have similar functions. In the business domain, clustering can be used to segment customers into groups for additional analysis and marketing activities. Clustering therefore relates to techniques from different disciplines including mathematics, statistics, computer science, artificial intelligence and databases. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents. It is also used in automatic document organization, topic extraction and fast information retrieval or filtering [20].

## 3. Difference Between Relative Terminologies

In this section we will show the main differences between classification, categorization and clustering in Table 1.1. These terminologies can be differentiated on the basis of motivation, process and model used. Another difference we have shown here is on the basis of algorithm used and application area. Each technique is associated with its own algorithm. Each technique can be used in different fields on its need and area. I have highlighted the main features associated with each terminology.

**Table 1.1:** Differences between Classification, Categorization and Clustering

| | CLASSIFICATION | CATEGORIZATION | CLUSTERING |
|---|---|---|---|
| **Motivation** | • Automatic classification for the large number of on-line text documents (Web pages, e-mails, corporate intranets etc.) | • Pre-given categories and labelled document examples (Categories may form hierarchy) | • Automatically group related documents based on their contents<br>• No predetermined training sets or taxonomies<br>• Generate a taxonomy at runtime |
| **Processes & Models used** | • Data pre-processing<br>• Definition of training set and test sets<br>• Creation of the classification model using the selected classification algorithm<br>• Classification model validation<br>• Classification of new/unknown text documents | • Automatic: Typically exploiting machine learning techniques<br>• Vector space model based<br>• Prototype-based (Rocchio)<br>• Neural Networks (learn non-linear classifier)<br>• Support Vector Machines (SVM)<br>• Probabilistic or generative model based | • Data pre-processing remove stop words, stem, feature extraction, lexical analysis etc.<br>• Hierarchical clustering-compute similarities applying clustering algorithms.<br>• Model-Based clustering (Neural Network Approach)- clusters are represented by exemplars (e.g.: SOM) |
| **Algorithm Used** | • Support Vector Machines<br>• K-Nearest Neighbours<br>• Naïve Bayes<br>• Neural Networks<br>• Decision Trees<br>• Association rule-based<br>• Boosting | • Naives Bayes, SVM<br>• K-Nearest Neighbour<br>• Decision Tree<br>• Neural Networking | • Sequential algorithms<br>• Hierarchical algorithms<br>• Agglomerative algorithms<br>• Divisive algorithms<br>• Fuzzy clustering algorithms |
| **Application** | • Document classification<br>• E-commerence interface (Amizon, ebay)<br>• Medical domain mesh<br>• Geo-demographic classification ACORN<br>• Data mining | • Web pages<br>• New articles/ events tracked & filtered by topic<br>• Journal articles index by subject categories<br>• Patents archived using international patent classification<br>• Email message filtering | • Document retrieval and texting<br>• Web support<br>• Pattern classification<br>• Image segmentation/ spatial data analysis<br>• Data mining (economic science, scientific data exploration and tools |

## 4. Comparison of Text Mining Techniques

In this section, main characteristic, algorithms, models and tools are shown in the Table 1.2. Text mining uses various numbers of techniques which play an important role. The techniques differ from each other. The information of retrieval technique used unstructured text where it can retrieve valuable information while as the information of extraction extracts the information from structured database. The Summarization technique is used to summarize the document which reduces length and keeps meaning same as it is.

The categorization is supervised process and uses predefined set documents according to their contents. Responsiveness and flexibility of the post-co-ordinate system effectively prohibit the establishment of meaningful relationships because a category is created by individual not the system. While as the clustering is used to find intrinsic structures in information and arrange them into related subgroups for further study and analysis. It is an unsupervised process through which objects are classified into groups called clusters. Clustering is dealing with high dimensional data, finding interesting pattern associated with data. Another feature is that it is a group of similar type of data and their relationship between them.

**Table 1.2:** Comparisons of Text Mining Techniques

| Techniques | Characteristics | Algorithms | Models | Tools |
|---|---|---|---|---|
| **Text Retrieval** | • Retrievals valuable information from unstructured text<br>• Document retrieval | • Stop Word Removal<br>• Stemming<br>• Word(term) extraction<br>• Inverted Index<br>• Signature file | • Boolean Model,<br>• Vector Space Model<br>• Statistical Language Model | • Intelligent Miner<br>• Text Analyst |
| **Text Extraction** | • Extract information from structured database.<br>• Feature retrieval | • Text location( Title/ Position)<br>• Cue phase in sentence<br>• Word frequencies (Text)<br>• Text cohesion link<br>• Discourse structure centrality<br>• Query-driven extraction | • HMM<br>• Conditional Markov Model<br>• SVM | • Text Finder<br>• Clear Forest Text |
| **Text Summarization** | • Reduce length by keeping its main points and overall meaning as it is | • Keyphase Extraction<br>• TextRank<br>• LexRank<br>• PageRank<br>• KEA<br>• ROUGE<br>• GRASSHOPPER | • Naïve Bayes Model | • Tropic Tracking Tool<br>• Sentence Ext Tool |
| **Text Categorization** | • Deal with large amount of text document<br>• Used in indexing documents to assist IR tasks as well as in classifying email or web pages in yahoo-like manner | • K-NN (K Nearest Neighbor Classification)<br>• Support Vector Machine<br>• Decision Tree Induction | • Support Vector Machines (SVM)<br>• Probabilistic or generative model based | • Intelligent Miner |
| **Text Clustering** | • Used as an unsupervised learning<br>• Goal is descriptive<br>• Cluster collection of documents<br>• Clustering, classification and sentimental analysis of text document | • K-Mean & K-Medoids<br>• Agglomerative & Divisive<br>• DBSCAN<br>• STING & CLIQUE | • Statistical Model<br>• Support Vector Machines (SVM) | • Carrot<br>• Rapid Miner |

## 5. Conclusion

In this review, the idea of text mining techniques have been introduced and presented. Due to its novelty, there are many potential research areas in the field of Text Mining, which includes finding better intermediate forms for representing the outputs of information extraction or retrieval. In this paper various text mining techniques are discussed with application. In addition to this we have compared the text mining technique on the basis of characteristic, algorithm used, models and tools.

## References

[1] Vishal Gupta and Guruprit Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 1, 2009.

[2] Shilpa Dang, Peerzada Hamid Ahmad, "Text Mining: Techniques and its Application", International Journal of Engineering & Technology Innovations, ISSN (Online): 2348-0866, Volume 1, Issue 4, pp. 22-25, 2014.

[3] Shah Neha K, "Introduction of Text mines and an Analysis of Text mining Techniques", PARIPEX, ISSN: 2250-1991, Volume 2, Issue-2, 2013.

[4] Amit Kumar Mondal and Dipak Kumar Maji, “Improved Algorithms for Keyword Extraction and Headline Generation from Unstructured Text”, First Journal publication from SIMPLE groups, CLEAR Journal, 2013.

[5] Peerzada Hamid Ahmad, Shilpa Dang, “A Comparative Study on Text Mining Techniques”, International Journal of Science and Research, ISSN: 2319-7064, Volume 3, Issue 12, pp. 2222-2226, 2014

[6] Rashmi Agrawal and Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Volume 2, Issue-6, 2013.

[7] Vidhya. K. A and G. Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 613-622, 2010.

[8] Amit Singhal, "Modern Information Retrieval: A Brief Overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp. 1-9, 2001.

[9] D. Subarani, "Concept Based Information Retrieval from Text Documents", IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661 Volume 2, Issue 4, pp. 38-48, 2012.

[10] Varsha C. Pande and A.S. Khandelwal "A Survey of Different Text Mining Techniques", IBMRD's Journal of Management & Research, ISSN: 2348-5922, Volume 3, No. 1, pp. 125-133, 2014.

[11] Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, pp. 141-160, 2003.

[12] Sayantani Ghosh, Sudipta Roy and Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Volume 1, Issue 4, pp. 223-233, 2012.

[13] Rahul Patel and Gaurav Sharma, "A survey on text mining techniques", International Journal of Engineering and Computer Science, ISSN: 2319-7242, Volume 3 Issue 5, pp. 5621-5625, 2014.

[14] Divya Nasa, "Text Mining Techniques- A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 2, Issue 4, pp. 50-54, 2012.

[15] Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Volume 2, No. 3, pp. 258-268, 2010.

[16] Oi Mean Foong, Alan Oxley and Suziah Sulaiman, "Challenges and Trends of Automatic Text Summarization", International Journal of Information and Telecommunication Technology IJITT, ISSN: 0976–5972, Volume 1, Issue 1, pp. 34-39, 2010.

[17] B. Mahalakshmi and K. Duraiswamy, "An Overview of Categorization Technique", International Journal of Modern Engineering Research (IJMER), ISSN: 2249-6645, Volume 2, Issue 5, pp. 3131-3137, 2012.

[18] S. Niharika, V. Sneha Latha and D. R. Lavanya, "A Survey on Text Categorization", International Journal of Computer Trends and Technology, ISSN: 2231-2803, Volume 3, Issue 1, pp. 39-45, 2012.

[19] Naveeta Mehta and Shilpa Dang, "A Review Of Clustering Techniques In Various Applications For Effective Data Mining", International Journal of Research in IT & Management, ISSN 2231-4334, Volume 1, Issue 2, pp. 50-66, 2011.

[20] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", Computer Science and Egineering, Volume 1, No 3, pp. 1-20.