International Journal of Advance Research in Science and Engineering
Vol. No.4, Issue 11, November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# A REVIEW ON VARIOUS TEXT MINING TECHNIQUES AND ALGORITHMS

**R. Balamurugan[1], Dr. S. Pushpa[2]**

*[1]Research Scholar, [2]Professor, Computer Science and Engineering,*
*St. Peter's University, Chennai (India)*

## ABSTRACT

*Text mining is the method of extracting meaningful information or knowledge or patterns from the available text documents from various sources. The pattern discovery from the text and document organization of document is a well-known problem in data mining. At present world, the amount of stored information has been enormously increasing day by day which is generally in the unstructured form and cannot be used for any processing to extract useful information, so different techniques such as classification, clustering and information extraction are available under the category of text mining. In order to find an efficient and effective technique for text categorization, various techniques of text categorization is recently developed. Some of them are supervised and some of them unsupervised manner of document arrangement. In this paper, focus is text mining process, different method of text categorization, cluster analysis for text documents, the basic differences between relative terminologies on the basis of process, model and the algorithms used, a comparison between text mining techniques on the basis of algorithms, models and tools used. In addition of that a new text mining technique is proposed for future implementation*

***Keywords: Classification, Clustering, Summarization, Techniques, and Algorithms***

## I. 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.

Text Mining 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. 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.

## II. RECENT STUDIES

This section of the paper explores recent efforts and contributions on text mining techniques. Therefore a number of research article and research papers and their contributions are placed in this section.

Many data mining techniques have been planned for mining valuable patterns in text documents. However, how to successfully use and update exposed patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the troubles of polysemy and synonymy. This paper presents an inventive and valuable pattern finding technique which includes the processes of pattern deploying and pattern evolving, to advance the effectiveness of using and updating discovered patterns for finding appropriate and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance [2].

The "helpfulness" characteristic of online user reviews helps consumers deal with information overloads and facilitates decision-making. However, many online user reviews require sufficient helpfulness votes for other users to assess their true helpfulness level. Text mining techniques are employed to remove semantic characteristics from review texts. Our findings also advise that reviews with strong opinions receive more kindness votes than those with mixed or neutral opinions. This paper sheds light on the considerate of online users helpfulness voting activities and the design of a enhanced helpfulness voting mechanism for online user review systems [3].

## III. TEXT MINING PROCESS

### 3.1 Document Gathering

In the first step, the text documents are collected which are present in different formats. The document might be in form of pdf, word, html doc, css etc.

### 3.2 Document Pre- Processing

In this process, the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step, the stages performed are as follows:

#### 3.2.1 Tokenization

The given document is considered as a string and identifying single word in document i.e. the given document string is divided into one unit or token

#### 3.2.2 Removal of Stop Word

In this step the removal of usual words like a, an, but, and, of, the etc. is done.

### 3.2.3 Stemming

A stem is a natural group of words with equal (or very similar) meaning. This method describes the base of particular word. Inflectional and derivational stemming are two types of method. One of the popular algorithm for stemming is porter's algorithm. e.g. if a document pertains word like resignation, resigned, resigns then it will be consider as resign after applying stemming method.

### 3.3 Text Transformation

A text document is collection of words (feature) and their occurrences. There are two important ways for representations of such documents are Vector Space Model and Bag of words.

3.4. Feature Selection (attribute selection):

This method results in giving low database space, minimal search technique by taking out irrelevant feature from input document. There are two methods in feature selection i.e. filtering and wrapping methods.

### 3.5 Data mining/Pattern Selection

In this stage the conventional data mining process combines with text mining process. Structured database uses classic data mining technique that resulted from previous stage.

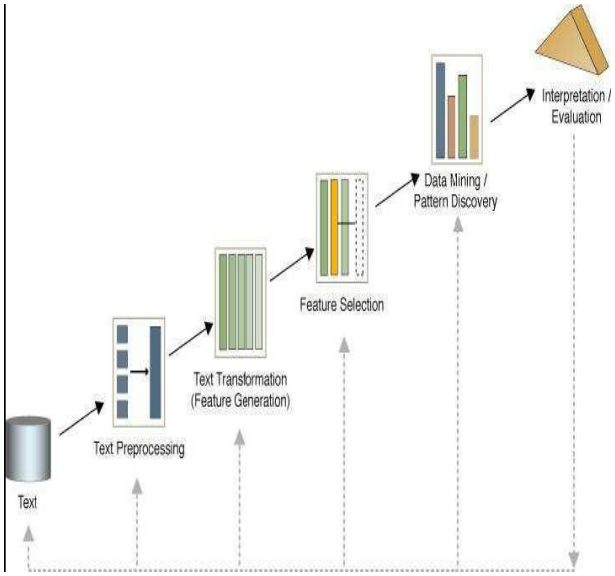


**Fig.1: Text Mining Process Flow**

### 3.6 Evaluate

This stage Measures the outcome. This resulted outcome can be put away or can be used for next set of sequence[2].

## IV. TECHNIQUES USED IN TEXT MINING

There are different kinds of techniques available by which the text pattern analysis and mining is performed. Some of the essential techniques are discussed in this section.

### 4.1 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 . 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. 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

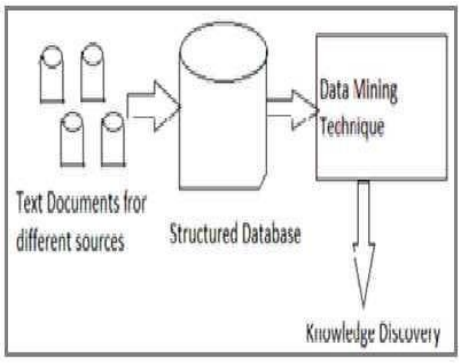


**Fig.2: Information Extraction**

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.

### 4.2 Topic Tracking

A topic tracking system apparatus by custody of user profiles and, based on the documents the user views, guess other documents of interest to the user. Yahoo offers free topic tracking tool that permits users to choose keywords and informs them when news relating to those topics becomes existing. Topic tracking methodology has its own limitations, however. For example, if a user sets up an alert for “text mining”, s/he will receive numerous news stories on mining for minerals, and very few that are really on text mining. Some of the improved text mining tools let users select specific categories of interest or the software routinely can even infer the user’s concern based on his/her reading history and click-through information

### 4.3 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].

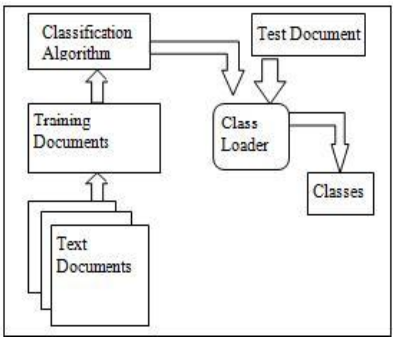


**Fig.3: Classification**

### 4.4 Text Clustering

Clustering [7] 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.

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.

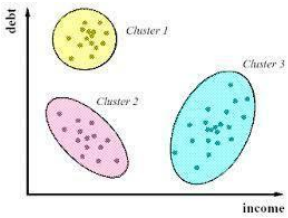


**Fig.4: Clustering**

### 4.5 Concept Linkage

Concept linkage tools [3]attach related documents by identifying their commonly-shared idea and help users find information that they perhaps wouldn't have establish using conventional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot. For example, a text mining software solution may

easily identify a link between topics X and Y, and Y and Z, which are familiar relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

### 4.6 Information Retrieval

Information retrieval[3] is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance. Due to the abundance of text information, information retrieval has found many applications

There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines. A typical information retrieval problem is to locate relevant documents in a document collection based on a users query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some adhoc information need, such as finding information to buy a used car. When a user has a long-term information need , a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user‟s information need. Such an information access process is called information filtering, and the corresponding systems are often called filtering systems or recommender systems. From a technical viewpoint, however, search and filtering share many common techniques. Below we briefly discuss the major techniques in information retrieval with a focus on search techniques.

### 4.7 Association Rule Mining

Association rule mining (ARM) [3] is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, specifically in education, counseling, and associated disciplines. ARM refers to the finding of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that often occur. Similar to the idea of correlation Study (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may contain two or more variables. This section provides the overview of text mining techniques and methodologies by which suitably text data becomes classifiable in next we discuss the data mining algorithms that are often consumed in the text mining and classification tasks.

### 4.8 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 . 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.

## V. TEXT MINING ALGORITHMS

There are various algorithms of data mining is available for effective classification and categorization. The discussion about whole methods and technique are not much feasible here therefore a little overview is proving in this section.

### 5.1 K Nearest Neighbour

In the text mining domain the k nearest neighbour algorithm is a classical and often used technique. In order to find a query text k nearest neighbour classifier is outperforms. This method estimates the distance between two strings for comparison and classify the text on the basis of distance.

$$d_A(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

Where x and y represents the data instances and d is distance between x and y. The main advantage of this algorithm is high accurate classification. On the other hand the major disadvantage is resources consumption such as memory and time.

### 5.2 Support Vector Machine

This approach is a one of most effective and accurate classification algorithm. In this approach concept using hyper-plans and dimension estimation based technique are used to discover or classify the data. The main advantage of this algorithm is to achieve high accurate classification results. But that is quite complex to implement.

### 5.3 Bayesian Classifier

That a probability based classification technique that is uses the word probability to classify the text data. In this classification scheme based on previous text and patterns data is evaluated and the class possibility is measured.

| Outlook | P | N | | Humidity | P | N |
|---|---|---|---|---|---|---|
| Sunny | 2/9 | 3/5 | | high | 3/9 | 4/5 |
| Overcast | 4/9 | 0 | | normal | 6/9 | 1/5 |
| Rain | 3/9 | 2/5 | | | | |
| Tempreature | | | | W indy | | |
| Hot | 2/9 | 2/5 | | True | 3/9 | 3/5 |
| mild | 4/9 | 2/5 | | False | 6/9 | 2/5 |
| cool | 3/9 | 1/5 | | | | |

That is some time slow learning classifier additionally that do not produces the more accurate results.

### 5.4 K-Mean Clustering

This technique is also a classical approach of text categorization. That uses the distance function as k nearest neighbour classifier to cluster data. That is an effective method of text mining in order to preserve the resources, but accuracy of this cluster approach is susceptible due to initial cluster center selection process. In addition of that hierarchical schemes of text categorization is available which are not much effective for cluster formation or categorization

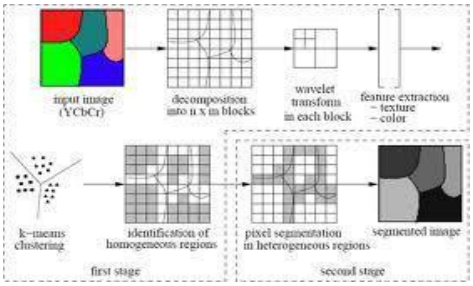


**Fig.5: K-Mean Clustering**

but comparative accuracy is much reliable than k-mean clustering.

## VI. DIFFERENCE BETWEEN RELATIVE TEMINOLOGIES

In this section we will show the main differences between classification, categorization and clustering. These terminologies can be differentiated on the basis of process and model used. Another difference we have shown here is on the basis of algorithm used. 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.

## 6.1 Algorithm Used

### 6.1.1 Classification

- Support Vector Machines
- Decision Trees
- K-Nearest Neighbours
- Naïve Bayes
- Neural Networks
- Association rule-based
- Boosting

### 6.1.2. Categorization

- Naives Bayes, SVM
- Neural networking
- Decision Tree
- K-Nearest Neighbour

### 6.1.3. Clustering

- Sequential algorithms
- Hierarchical algorithms
- Agglomerative algorithms
- Divisive algorithms
- Fuzzy clustering algorithms

## 6.2 Processes and Models Used

### 6.2.1. Classification

- Data pre-processing
- Definition of training set and test sets
- Creation of the classification model using the selected classification algorithm
- Classification model validation
- Classification of new/unknown text documents.

### 6.2.2. Categorization

- Automatic: Typically exploiting machine learning techniques
- Vector space model based
- Prototype-based (Rocchio)
- Neural Networks ( learn non-linear classifier)
- Support Vector Machines(SVM)
- Probabilistic or generative model based

### 6.2.3. Clustering

- Data pre-processing, remove stop words, stem, feature extraction, lexical analysis etc.,
- Hierarchical clustering-compute similarities applying clustering algorithms.

- Model-Based clustering(Neural Network Approach) – clusters are represented by exemplars(e.g: SOM)

## VI. COMPARISON OF TEXT MINING TECHNIQUES

In this section, main algorithms, models and tools are shown. Text mining uses various numbers of techniques which play an important role. The techniques differ from each other. 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 I 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.

### 7.1 Technique – Summarization:

#### 7.1.1. Algorithms

- Keyphase Extraction
- TextRank
- LexRank
- PageRank
- KEA
- ROUGE
- GRASSHOPPER

#### 7.1.2. Model

- Naïve Bayes Model

#### 7.1.3. Tools

- Tropic Tracking Tool
- Sentence Ext Tool

### 7.2 Technique – Categorization:

#### 7.2.1 Algorithms

- K-NN (K Nearest Neighbor Classification)
- Support Vector Machine
- Decision Tree Induction

#### 7.2.2 Models

- Support Vector Machines (SVM)
- Probabilistic or generative model based

#### 7.2.3 Tool

- Intelligent Miner

### 7.3 Technique – Clustering

#### 7.3.1. Algorithms

- K-Mean & K-Medoids
- Agglomerative & Divisive
- DBSCAN
- STING & CLIQUE

#### 7.3.2. Models

- Statistical Model
- Support Vector Machines (SVM)

#### 7.3.3. Tools

- Carrot
- Rapid Miner

## VIII. CONCLUSIONS

In this paper various techniques and methods are discussed for effective and accurate text mining. In addition of that the effective algorithms are also learned. Due to observation a promising approach is obtained given in [5]. According to the analyzed methods an improvement over the [5] is suggested. In near future the proposed technique is implemented using Rapidminer and MATLAB tool and the comparative results are provided.

## IX. ACKNOWLEDGMENT

I wish to thanks who directly and indirectly contribute in paper, first and foremost, I would like to thank Prof. Dr. S. Pushpa for her support and encouragement. she kindly read my paper and offered valuable details and provide guidelines. Second, I would like to thanks all the authors whose paper i refer for their direct and indirect support to complete my work.

## REFERENCES

[1] Vishal Gupt a, Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of emerging technologies in web intelligence Vol. 1, No.1, August 2009.

[2] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, “Effective P attern Discovery for Text Mining”, IEEE Transact ions on Knowledge and Data Engineering. C Copyright 2010 IEEE

[3] Qing Cao, Wenjing Duan, Qiwei Gan, “Exploring determinant s of voting for t he “helpfulness” of online user reviews: A text mining approach’, 0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009

[4] Hamid Mousavi, Shi Gao, Carlo Zaniolo, “ IBminer: A Text Mining Tool for Constructing and Populating InfoBox Databases and Knowledge Bases”, Proceedings of the VLDB Endowment , Vol. 6, No. 12, Copyright 2013 VLDB Endowment 21508097/13/10...$ 10.00.

[5] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky, "HierarchicalTopics: Visually Exploring Large Text Collect ions Using Topic Hierarchies",IEEE Transaction on Visualization and Computer Graphics, Vol. 19, No.12, December 2013

[6] Liwei Wei, Bo Wei, Bin Wang, "Text Classification Using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58, doi:10.4236/jsea.2012.512b012 Published Online December 2012

[7] Lokesh Kumar and Parul Kalra Bhatia,"TextMining:Concept Process, Applications," Journal of Global Research in Computer Science Volume 4, No. 3, March 2013 .

[8] P. Monali , K. Sandip, "A Concise Survey on Text Data Mining" in proceeding of the International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014, pp 8040- 8043.

[9] Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009