International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
1st International Conference on Advent Trends in Engineering, Science and Technology
"ICATEST 2015", 08 March 2015

A Survey on Text Mining- techniques and application

Ruchika R. Tated1, Mangesh M. Ghonge2

1Department of Computer science & Engineering, JCOET Yavatmal, India. ruchikatated@gmail.com,
2Professor, Department of Computer Engineering, JCOET Yavatmal, India. mangesh.cse@gmail.com

Abstract— Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. The pattern discovery from the text and document organization of document is a well-known problem in data mining. In today's 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 several 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 on concept of text mining, text mining process, techniques used in text mining also presenting some real world applications of text mining. In addition, brief discussion of text mining benefits and limitations has been presented.

Index Term- Text mining, Techniques, Classification, Clustering, Information Extractions, and Applications.

1.INTRODUCTION

The amount of stored information has been enormously increasing day by day, so discovering patterns and trends out of massive data is a great challenge. In this work, a discussion over the techniques which can be used to resolve this problem is done. The main technique is data mining, the application of which are text mining and web mining. The focus of this work is text mining which is explained in the following sections.

Text is the most common vehicle for the formal exchange of information. Although extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost.

Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [13]. The aim of text mining tools is to be able to answer sophisticated questions and perform text searches with an element of intelligence.

Technically, text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents a step forward from text retrieval. It is a relatively new and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. Text mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from text. Researchers like [15], [14] and others pointed that text mining is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT).

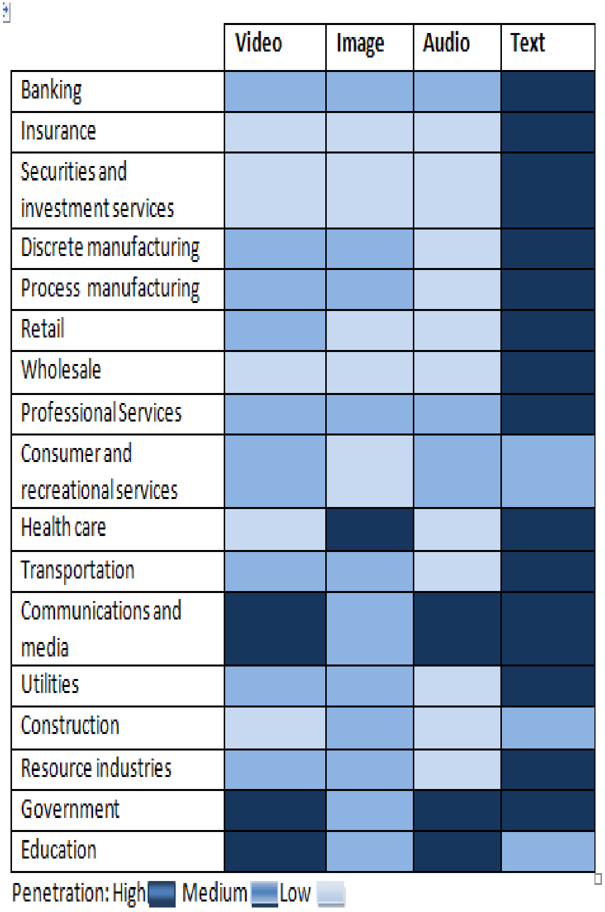
According to [10] the boundaries between data mining and text mining are fuzzy. The difference between regular data mining and text mining is that in text mining, the patterns are extracted from natural language texts rather than from structured databases of facts. The goal of data mining is to discover the implicit, previously unknown trend and patterns from the databases. Data mining consists of many techniques such as classification, clustering, neural networks, and decisions trees.

2. 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 discovery 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 propose [21].

Fig. 1 shows that, maximum amount of data used by various sectors is in textual form. Text mining is one of the important fields of data mining dealing with unstructured or semi-structure data. Text mining introduces enumerate model from linked research domain like classification, clustering etc. Numerical measures can be deriving by enforcing Text analysis methods to unstructured textual information [1].



	Video	Image	Audio	Text
Banking	▦	▦	▦	■
Insurance	□	□	□	■
Securities and investment services	□	□	□	■
Discrete manufacturing	▦	▦	□	■
Process manufacturing	▦	▦	□	■
Retail	▦	□	□	■
Wholesale	□	□	□	■
Professional Services	▦	▦	▦	■
Consumer and recreational services	▦	□	▦	▦
Health care	□	■	□	■
Transportation	▦	▦	□	■
Communications and media	■	▦	■	■
Utilities	▦	▦	□	■
Construction	□	▦	□	▦
Resource industries	▦	▦	□	■
Government	■	▦	■	■
Education	■	▦	■	▦

Penetration: High ■ Medium ▦ Low □

Fig.1. Types of data available and generated by various sectors.

3. TEXT MINING PROCESS

1) Document Gathering:

In the first step, the text documents are collected which are present in different formats[6]. The document might be in form of pdf, word, html doc, css etc.

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 [6][16]:

a) 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[6].

b) Removal of Stop word:

In this step the removal of usual words like a, an, but, and, of, the etc. is done [12].

c) 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[10]. One of the popular algorithm for stemming is porter's algorithm[11]. e.g. if a document pertains word like resignation, resigned, resigns then it will be consider as resign after applying stemming method[12].

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 [2].

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.

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 [2].

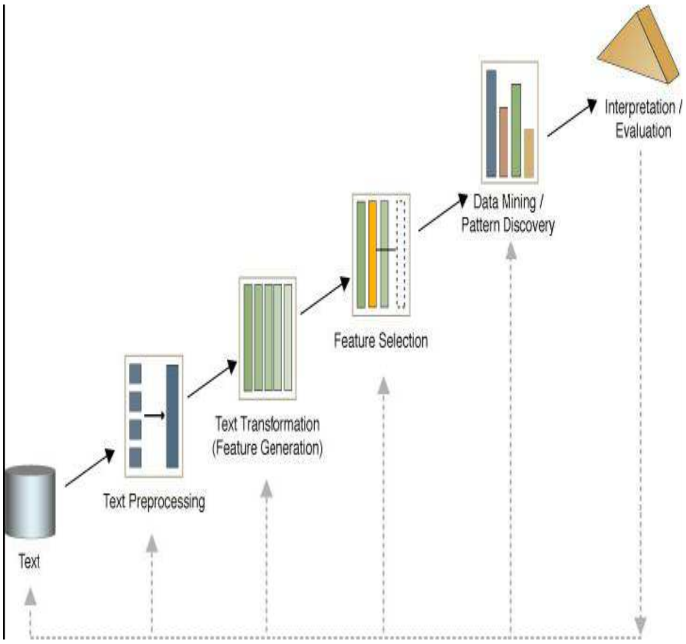

Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

Fig.2. Text Mining Process flow.

6) Evaluate:

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

4. TEXT MINING TECHNIQUES

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 Categorization:

It is a supervised technique. A supervised technique is one which is based upon the set of input-output examples which are basically used to train the model being used, in order to classify the new documents. Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content [14].

It is process of finding main theme of document by adding metadata and analyzing document [17]. This technique find counts of words and from that count decides topic of the document. In this process, text documents are classified into predefined class label [8].

A number of classification techniques can be applied to categorize the text, here decision tree classifier are explained.

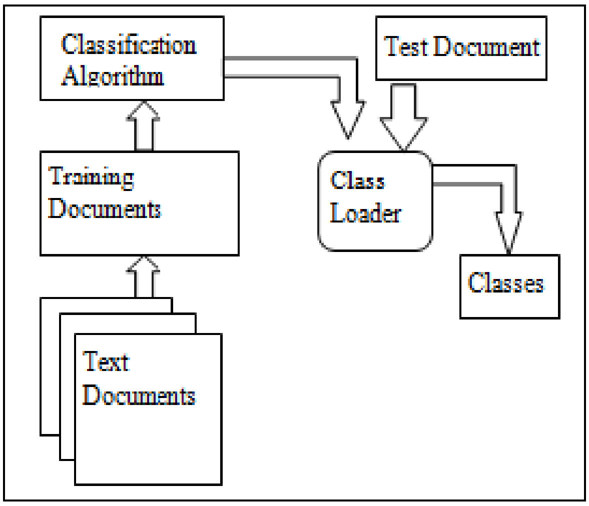


Fig. 3: Classification

4.1.1 Decision Trees

Decision tree methods rebuild the manual categorization of the training documents by constructing well-defined true/false queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After having created the tree, a new document can easily be categorized by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf. The main advantage of decision trees is the fact that the output tree is easy to interpret even for persons who are not familiar with the details of the model [18]. The tree structure generated by the model provides the user with a consolidated view of the categorization logic and is therefore useful information. A risk of the application of tree methods is known as "over fitting": A tree over fits the training data if there exists an alternative tree that categorizes the training data worse but would categorize the documents to be categorized later better.

This circumstance is the result of the algorithm's intention to construct a tree that categorizes every training document correctly; however, this tree may not be necessarily well suited for other documents. This problem is typically moderated by using a validation data set for which the tree has to perform in a similar way as on the set of training data. Other techniques to prevent the algorithm from building huge trees (that anyway only map the training data correctly) are to set parameters like the maximum depth of the tree or the minimum number of observations in a leaf. If this is done, Decision Trees show very good performance even for categorization problems with a very large number of entries in the dictionary.

4.2 Clustering:

Text Clustering is an unsupervised technique in which no input out patterns are pre - defined. This method is based upon the concept of dividing the similar text into the same cluster. Each cluster consists of number of documents. The clustering is considered better if the contents of documents of intra cluster are more similar than the contents of inter-cluster documents.

Clustering is a technique used to group similar documents but it differs from classification in than documents are clusters on the fly instead of through the use of pre-defined topics [8]. After calculating similarity, clustering algorithms [3] are applied to generate list of classes. This process of clustering is depicted in Figure 4.

Clustering can be divided into following categories: hierarchical clustering and partitional clustering.

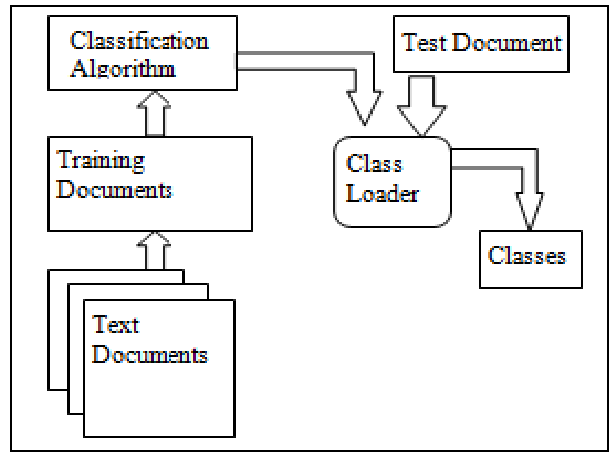


Fig. 4: Clustering

4.2.1 Hierarchical Clustering:

This clustering uses the cosine similarity measure. The result of hierarchical clustering is a single clustered tree. It works at different level of granularity. It can be divided into two categories: i) Bottom up hierarchical clustering method ii) Top down hierarchical clustering method.

i) Bottom up hierarchical clustering method:

Every document is considered as a separate cluster, and then on the basis of similarity, clusters are combined repeatedly till single cluster is formed.

The steps can be summarized as follows:

1) Consider each document as a single cluster.
2) Calculate similarity of cluster ai with cluster bj then merge the two having maximum similarity.
3) Repeat step 2 till single cluster is formed.

ii) Top down hierarchical clustering method:

In this method, work starts from a single cluster as whole, then it get split iteratively into various clusters on the basis of smallest similarity measure. Top down approach does not have much application as compare to bottom up approach. This is much complex as compare to bottom up approach as amount of computations involved are quite large.

The advantage of hierarchical clustering is that: it can use any form of similarity measure and the disadvantage is that once the clusters are formed, cannot be rebuilt, to improve performance, if needed.

4.3 Information Extractions:

Natural language text documents contain information that cannot be used for mining. As documents are considered as ―bag of words‖ they can be represented by vector model which then can be used as an input to the above defined techniques such as classifications, clustering but this is not used for this method. In Information extraction, the documents are first converted into the structured databases on which data mining techniques can be applied to extract knowledge or interesting patterns. The following Fig. 5 shows the process[5].

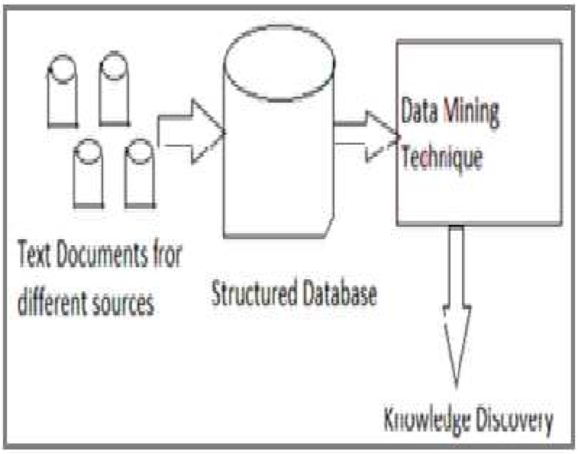


Fig. 5: Information Extraction

The task of IE is the identifications of entities, i.e. person names, location name, company name etc. The required pieces of information such as ―position‖, ―person name‖ are found. The outcome would be a template in which all the entities and their relationships with one another can be easily identified. Then the information is entered into the database so that data mining techniques can be applied in order to find some implicit information.

5. APPLICATIONS

5.1 Business Intelligence:

Text mining techniques helps for determining particular topic or event as in business decision support system amount of cutting down the cost of predicting future work is an important task [8].

5.2 Bioinformatics:

Nowadays, biomedical articles have occupied major area in different applications [4]. The aim of text mining in bioinformatics is to allow researchers to explore advance knowledge in the field of biomedical in an efficient way.

5.3 Security Application:

In network security for encryption and decryption techniques, text mining methodologies are used with the intention of security at national level application, the analysis and monitoring of documents like plain texts, emails, web blog articles text mining is used [12].

5.4 Human Resource Management:

For purpose of recruiting the candidate by reading and writing their CVs text mining is used [12]. Also for growth of particular organization for the application prefers the monitoring of level of customer, examining staff's satisfaction text mining techniques applied.

5.5Web Search Enhancement:
In text mining, by using *text categorization* techniques such as CatS[20]. The presentation of result is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered "maps" (kartoo.com) [9].

5.6 Customer Relationship Management (CRM):
Text mining is also useful in Customer Relationship Management (CRM) for supplying immediate answers to frequently asked questions

5.7 Text mining is also used in following sectors [7]-
i) Publishing and media.
ii) Telecommunications, energy and other services industries.
iii) Information technology sector and Internet.
iv) Banks, insurance and financial markets.
v) Pharmaceutical and research companies and healthcare.

6. CONCLUSION

Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Here in this work quite big research field ―Text Mining‖ is discussed with its various techniques which can be used such Classification, a supervised technique i.e. having all the input output patterns which are used to train the model, before it can be used to classify the newly arrived document. Clustering is used to divide the text into the clusters according to the similarity of the documents. It is an unsupervised learning technique in which, no pre-defined input-output patterns are there. Information Extraction is basically used to extract structured information from the unstructured text, on which data mining techniques can be applied for getting useful patterns or knowledge from the documents. Applications in the field such as identifying business intelligence, bioinformatics, security application, web search enhancement, CRM and other sectors are studied.

REFERENCES

[1]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.*

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

[3]Falguni N. Patel, Neha R. Soni," Text mining: A Brief survey", International Journal of Advanced Computer Research (ISSN (pri nt): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.

[4] Of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[5] 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, April 2012 pp 51-540, in IJARCSSE

[6]Vandana Korde and C. Namrata Mahender, "Text Classification and Classifiers :A Survey", International Journal of Artificial Intelligence & Application, Vol.3, No.2, March 2012.

[7]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

[8] V. Gupta, G.S. Lehal ― A Survey of Text Mining Techniques and applications ―, Journal of Emerging Technologies in Web Intelligence,2009. Issue.

[9] Novi Sad J. Math, Milos Radovanovic and Mirjana Ivanovic," Text Mining: Approaches and Applications" Vol. 38, No. 3, 2008, 227-234.

[10] R. Malik, "Conan: Text mining in biomedical domain," PhD thesis, Utrecht University, Austria, 2006.

[11] Andreas Hotho, Andreas Nurnberger and Gerhard PaaB, "A Brief Survey of Text Mining", May 13, 2005

[12] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz,"A decision-tree-based symbolic rule induction system for text categorization", IBM Systems Journal, September 2002.

[13] F. Sebastiani, "Machine learning," *ACM Computing Surveys*, vol. 1, no. 34, pp. 1–47, 2002.

[14] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 3–10.

[15] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)," in *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 1995, pp. 112–117.

[16] R. Sagayam, S. Srinivasan and S. Roshni", A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques," International Journal Of Computational Engineering Research Vol. 2 Issue.

[17]http://www.planetdata.eu/sites/default/files/presentations/Big_Data _Tutorial_part4.pdf
[18]http://www.isical.ac.in/~acmsc/TMW2014/M_mitr a.pdf
[19] http://www.abbottanalytics.com
[20]Jonathan G. Fiscus and George R. Doddington, " Topic Detection and Tracking Evaluation Overview".
[21] Charu C Aggrawal and Chengxiang Zhai,"Mining Text Data".