Volume 2, Issue 4, April 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Text Mining Techniques- A Survey

Divya Nasa

USICT , GGSIPU

divya.nasa.jul@gmail.com

Abstract— **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 summarization, classification, clustering, information extraction and visualization are available for the same which comes under the category of text mining. Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents. In this work, a discussion over framework of text mining with the techniques as above with their pros & cons and also applications of Text Mining is done. In addition, brief discussion of Text Mining benefits and limitations has been presented.**

Keywords— **Text mining Framework, Techniques, summarization, classification, clustering, information extraction, visualization, applications, benefits and limitations.**

I. 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.

The basic form of information is data which is to be managed and mined in order to create the knowledge. Data mining emerged in the 1980's to resolve the above problem [1]. 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. Data mining is process which requires data pre-processing before applying any technique. It's a part of knowledge discovering process[KDD]. The main steps of KDD[2] are :

- In the first phase, business objectives and expectation are defined.
- In the second phase, the above defined objective helps in the selection of data from the data ware house. Then that data is pre-processed in order to improve quality of data.
- In the third phase, the algorithm of data mining is selected and applied to the data which was prepared in second step. This is a vital step, of knowledge discovery process. The relationship and patterns would be the output of this phase.
- In the fourth phase, analysis of the relationship & patterns is done and valid patterns according to the objectives are found out.
- The last phase is the visual representation of the knowledge discovered. These results can be stored, assembled which can be used to improve the business.

The applications of data mining are Web mining and Text mining.

"Web mining [3] is basically deriving knowledge or interesting information from the Web data. It is normally expected that either web log data or hyperlinks structure of the web or both have been used in the mining process. Web mining can be divided into categories web context mining, web structure mining and web usage mining. These can be referred from [3].Text mining is an application of Data mining. The basic difference between the two is that in text mining patterns are extracted from natural languages text rather than from structured databases unlike data mining. The text mining study has been shown in the following sections.

II. TEXT MINING FRAMEWORK

Definition: Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. As the text is in unstructured form, it is quite difficult to deal with it. Finding "nuggets" of interesting information from the natural language text is the purpose of text mining.

The Text Mining Process is shown in Fig. 1:g

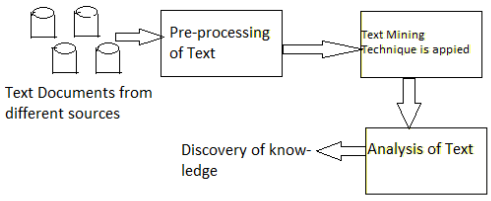

Fig. 1 Text Mining Process

Stage-I: Pre-processing Text:

Mining from a pre-processed text is easy as compare to natural languages documents. So, pre-processing of

documents that are from different sources is an important task during text mining process before applying any text mining technique.

As Text documents can be represented as "bag of words" on which different text mining methods are based. Let Ω be the set of documents & W= {w_1, w_2, ---- w_m} be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words, which do not provide relevant information; stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed that contain no informatics as such. Stemming methods: are used to produce the root from the plural or the verbs. For e.g. Doing, Done, Did may be represented as 'Do'. After this method is applied, every word is represented by its root word. Originally it was proposed by M. poster [4].

Stage II- Text Mining Technique is applied:
This is an important stage in which the selected algorithm is applied on text in order to process the text. The algorithm such as clustering, classification, summarization, information extractions or visualizations which are explained next could be used.

Stage III - Analysis of Text:
Here the outputs are analysed for discovering the knowledge. Various tools such as link discovery tool can be used or the outputs can be visualised so that the users could navigate through in order to achieve the perspective.

III. TECHNIQUES OF TEXT MINING

Summarization of Text:

Due to great amount of information, there is a need of producing summaries from number of documents. In this technique, length of the document is reduced such that meaning and main points should not be lost. As summary can be produced from a single document or group of documents. A summary can replace the set of documents. A summary contains a significant position of information in the original document(s) and that is no longer than half of the original documents[5].

The process of summarization can be divided as:-

Pre-processing: In this step, document(s) are represented in the structured form, by reducing the dimensionally of the documents. This can be done by applying filtering methods, stemming methods as explained in 3.1. Also, as each sentence is a combination of words, is represented by the using vector model, in which each sentence is considered as an N-dimensional vector. It makes possible in finding the similarity between the different text elements. To find similarity of the two vectors, cosine of the angle of the two vectors of document set is found out using the formula:

$$\text{Cos } \theta = \frac{\vec{x}\vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

Where x & y are the 2 vectors. If Cos θ = 1 it corresponds to maximum similarity, if Cos θ = 0, it corresponds to minimum similarity. The performance degrades by using vector model if the dimensionality of documents is quite high. To overcome this problem, an approach called random Indexing [6] was developed. Here every word is assigned a random vector called as index vector. The advantage of Random indexing over vector model is that similarity computations can be performed even after less number of vectors has been encountered unlike vector model which require entire data to be scanned before computations can be performed.

Categorisation:

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. In this method, pre-defined classes are assigned to the text documents. The goal is to train the classifier on the basis of known examples and then unknown examples are categorized automatically. Here for reducing the dimensionality of the document set, a method called as Index Term Selection [7] is used. In this method, for every word information gain[7] is found out which can be as follows :

$$IG(t_j) = \sum_{c=1}^{2} p(L_c) \log_2 \frac{1}{p(L_c)} - \sum_{m=0}^{1} p(t_j=m) \sum_{c=1}^{2} p(L_c|t_j=m) \log_2 \frac{1}{p(L_c|t_j=m)}$$

Here $p(L_c)$ is the fraction of training documents with classes L_1 & L_2 , $P(t_j=1)$ and $p(t_j=0)$ is the number of documents with or without terms t_j and $p(L_c/\ t_j\ =m)$ is the conditional probability of classes L_1 and L_2 if term t_j is contained in the document or is missing. For every word IG is found out. The words with very low information gain as compare to some threshold value are removed.

A number of statistical classification techniques can be applied to categorize the text, here Naïve Bayesian Classifier, and Nearest Neighbour classifier are explained.

Naïve Bayes Classifier:

It depends upon the probabilistic relationship between different categories. If this classifier is considered as a graph, then there is a parent node connected to the child nodes, no other connection would be possible. Assume that each instance of document is represented as a vector Y= {w_1 ,w_2 , ---- w_m } where 'w' represents the words in the document and the classes are represented as C={c_1, c_2,-------c_k}. If category of new instance of Y is to be found out then, its probability is calculated with all the classes. It is put in the category, having the maximum probability. The Bayesian formula for calculating the conditional probability of instance of Y to which class it belongs is :

$$P(C_i \mid Y) = \frac{P(Y \mid C_i)\, P(C_i)}{P(Y)} = \frac{P(Y \mid C_i)\, P(C_i)}{\sum_{J=1}^{k} P(C_j)\, P(Y|C_j)}$$

This classifier assumes that attributes are totally independent to one another then,

$$P(Y|Ci) = \Pi_{j=1}^{m} P(wj|Ci)$$

Naïve Bayesian is simple and efficient to implement as it assumes that all the words of the documents are independent to one another. The authors [7] used a combination of Expectation - maximization & a Naïve bayes classifier and were able to reduce the classification error by up to 30%.

Nearest Neighbour Classifiers:

In this type of classifier, similarity of the unknown document is calculated with all other document in the training set, if k similar documents are considered then, it is called as k nearest neighbour classifier. The similarity between the documents is found out using the cosine similarity formula as explained in 3.1.

Let us assume if document $d_n \varepsilon D$ and $d_n \varepsilon C_k$ class. Then to find the class of the documents d_m to which it belongs. Similarly measure $S(d_n, d_m)$ is calculated with respect to all $d \varepsilon D$. The unknown document d_m is put into the class of the document d_n with highest probability. The disadvantage of this method is the overhead of calculating similarly measure w.r.t. all the document in the training set.

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 categorization in than documents are clusters on the fly instead of through the use of pre-defined topics [8].Clustering can be divided into following categories: hierarchical clustering and partitional clustering.

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.

Bottom up hierarchical clustering method:

Every document is considered as a separate cluster, 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 a_i with cluster b_j then merge the two having maximum similarity.
3) Repeat step 2 till single cluster is formed.

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.

Patitional Clustering:

The task of text partitional clustering [9] is that document set is divided into k disjoint point sets. In this method, a function is chosen such that distance of every point and the centroid of the cluster it belongs to, should be minimum. k- means algorithm is a partitional clustering algorithm which is explained as follows.

k-means Algorithm:

This method can be applied to large data sets. The aim of the method is that k clusters are formed from the data set. Recursive updating of centroids is done in this method. Each cluster would have a reference point known as centroid which will be used in every round of iterations. It works as follows: Assume we have document set D ($d_1, d_2 \ldots d_m$).

(1) k– data points are chosen as initial centroids.
(2) Calculate distance of each $d \varepsilon D$ to the centroids, assign it to the closest centroid.
(3) Centroid is recomputed until centroid become stable.

Its advantage is that its performance is good when applied to convex clusters. It is the most frequently used algorithm in the field of text mining. The main disadvantage is that everything is dependent on the centroid, there is no scope of scalability and it cannot be applied to non-spheres clusters.

Hierarchical Clustering method becomes problematic for large data sets as the memory required to store the similarity matrix, which consists of n(n-1)/2 elements where n is the number of documents and also run tine behaviour with $O(n^2)$ is worse as compared to the linear behaviour of k means algorithm[6].

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. 2 shows the process:

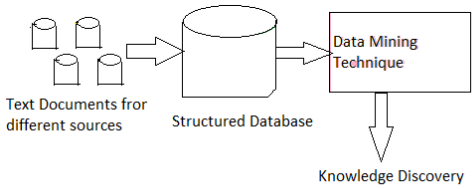

Fig. 2 Information Extraction

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.

Visualization:

Visualisation method provides better and faster understandable information which helps us to mine large documents collections. Unlike the descriptions which are purely text base. By using this method, the users can distinguish between colors, relationships, distance, etc. Thousands of points can be scanned easily via this model. This rely on the fact of presenting the discoveries in the form of graphs, maps etc. which makes users to understand that quickly. "The eye, the hand and the mind seem to work smoothly and rapidly as users perform actions on visual displays" [10] Visualization for the documents can be shown via a diagrammatic representation in Fig. 3 as below.

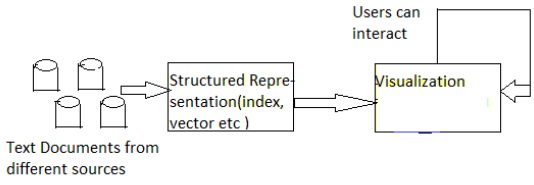

Fig. 3 Visualization

The collection of documents can be represented is a structured format using indexing or vector space model etc. then the visualization technique can be applied . The users can interact directly to refine their needs. Refinement while navigating through the maps/graphs makes users possible to achieve the goal. Now a days, in text mining, visualization methods can improve the method for extracting the interesting patterns as it provides the visual representation which simplifies the task.

For visualizing document collections, the high dimensionally of the document set is represented usually by 2-D representations. The colours between the documents represent the distance between the neighbours. If the colours are dark than the documents are much similar as compare to if they are light. Visualization can be done using the neural network technique called SOM (Self organizing maps). It is an unsupervised technique which is used to convert the high dimensionally of documents into usually two dimensional representations. Advantage of this technique is that it makes quick detection of patterns or trends in large data set. And also it can handle documents with high dimensionality.

IV. APPLICATIONS AND MERITS/DEMERITS

• Classification of news as a Text : As number of stories are there in a daily newspaper, the users would like to see the stories in which names of different persons are involved for some place, organization etc. Manually doing such a task is a tedious job, text mining method i.e. Information extraction can be used to perform such a task, which would retrieve templates containing different entities and their relationship with one another in the structured format, which would be put in the databases , on which Data Mining techniques can be applied for retrieving the interesting patterns.

• Analysis of the Market trends: There is a need, to know the market conditions for the growth of an organization, such as its number of competitors, number of employees, sales, products etc. of the competing organizations. For such analysis enormous amount of information is there, doing all such analysis manually is near to impossible. Earlier there was a separate division in every organization for such kind of work, but due to the arrival of Text Mining Techniques, it becomes simple to handle. Either classification technique can be used or Information Extraction for simplifying the task.

•Analysis of the junk Emails : Another common application for text mining is in automatic analysis of the junk E-mails which are undesirable. Classification technique of text mining can be used to classify such mails on the basis of pre-defined frequently occurring terms.

Demerits of Text mining:

i) No programs can be made in order to analyse the unstructured text directly, to mine the text for information or knowledge.

ii) The information which is initially needed is no where written.

Merits of Text mining:

i) As database can store less amount of information, this problem has been solved through Text Mining.

ii) Using the technique such as information extraction, the names of different entities, relationship between them can easily be found from the corpus of documents set.

iii) Text mining has solved the problem of managing such a great amount of unstructured information for extracting patterns easily, otherwise it would have been a great challenge.

V. CONCLUSIONS

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 as summarization, which is basically used to produce the relevant information from the corpus. 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. Graphical Visualization is used to provide better understandable information for mining the documents. Applications in the field such as identifying news stories, junk emails , and analysis of the market is studied.

REFERENCES

[1] A Lew and H. Mauch " Intoduction to Data Mining Principles" ,SCI, springer, 2006.
[2] P. Ponniah "Data Warehousing Fundamentals.
[3] N. Verma, U. ghose "An Overview of Web Content Mining with Cleaning of Redundant Links from Web Pages ",2006
[4] M. Porter " An Algorithm for Suffix Stripping".
[5] Hovy, E.H. "Automated Text Summarization", 2005.
[6] A. Hotho, A .N. berger, G. Paab . "A Brief Survey of Text Mining", 2005.
[7] W. D. S. Yu, Q.W. J. Yu and Q. Guo, " A Novel Naive Bayesian Text Classifier".
[8] V. Gupta, G.S. Lehal " A Survey of Text Mining Techniques and applications ", Journal of Emerging Technologies in Web Intelligence,2009.
[9] F.Liu , Lu Xiong "Survey on Text Clustering Algorithm".
[10] E. Morse. "Document Visualization".

