Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Survey on Text Mining Procedures and Exploring Techniques

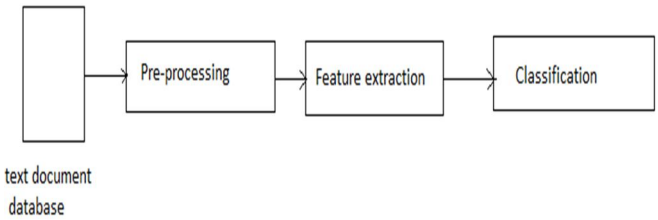
V. Juli Stephy [1], C.Pabitha [2]
[1]P.G. Student, [2]Assistant Professor , Department of Computer Engineering
Valliammai Engineering College, Chennai, India.

Abstract: Amount of data are increased in today world. we can extract the useful information which is generally in the unstructured form. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analysed with traditional data mining techniques More number of techniques available such as information extraction, clustering, classification, summarization, visualization are available under the text mining techniques. In this paper discuss about the text mining and techniques and text mining benefits and limitation has been presented.

Keywords: ***Text mining; Information extraction; clustering; classification; summarization***

I. INTRODUCTION

Text mining is extract the meaningful information from the text. It is use the data mining algorithm. Text mining is also called as text data mining. It derives the high quality of information from text. Text mining divides the pattern with in structure and process the pattern and evaluates it. Finally it produces the output. Text mining is like a text data mining which one is applied on textual data. It extracts the information from the unstructured data. Keyword search method is used in text mining. It contains pre-defined keyword. Keyword such as entities phrases or sentences. Text mining is based on the natural language process (NLP). It is used to read and analyse the textual information. Natural language process based on the queries. These queries are the Pre-written queries. In this text mining the pattern are extracted from the unstructured data or natural language text. In data mining the pattern are extracted from the database. In text mining the input is unstructured text. In web mining the input is structured. In sentence spitting Identifying sentence boundaries in a document is not as trivial a process as it may seem. SEASR has components that achieve sentence splitting either using rules or statistical models (or both). Once sentences are identified they are recorded as annotations in their own annotation set. Tokenization, simply put, is basically labelling individual words or sometimes word parts. This is important because many down stream components need the tokens to be clearly identified of analysis. Tokens are recorded as annotations in their own annotation set. Part of speech Such components typically assign a POS tag to a token (the Penn Treebank project has provided a set of codes for this purpose that is widely used). Other data such as lemma, lexemes, and synonyms (to name a very few) may also be identified at this stage. POS information is stored as features of the token annotation. Text mining contain five steps, Collection of text document, pre-processing of text, Text mining techniques, analyse the text, knowledge discovery.



First collect the text documents. The pre-processed text is easy to compare with the natural language text .so we can compare the pre-processed text with the natural language text. In this phase two method are involved. Filtering and streaming . Filtering is used to remove the unwanted words (or) which one is do not provide the relevant information. Streaming words provide the root for the related words. After applying the streaming method every word is represented by it's root node.

II. LITRATURE REVIEW

Paper presented by Ning Zhong, Yuefeng Li, and Sheng-Tang [1] that presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and

TREC topics demonstrate that the proposed solution achieves encouraging performance.

Paper presented by Charu C. Aggarwal, Philip S. Yu demonstrates [2] that segmentation of text data records is required in many application such as document organization, new group filtering, text crawling. The categorical data stream clustering problem of customer segmentation. Statistical summarization methodology an online approach for clustering massive text and categorical data streams is presented here.

Paper presented by Douglass Michael Steinbach George Karypis Vipin Kumar demonstrates [3] that two main to document clustering, agglomerative hierarchical clustering and k-means are compared here. Hierarchical clustering is always better quality clustering approach. K means have the time complexity. Sometime k means and agglomerative are merged and get the best of both world. K means technology is better than the k means approach as good or better than the hierarchical approach. An explanation for these results that is based on analysis of the specific clustering algorithm and the nature of document data is proposed here.

Paper presented by S. Zhong demonstrates [4] that clustering data streams has been a new research topic, recently used in many real data mining applications, and has attracted a lot of research attention. However, there is not much work on clustering high-dimensional streaming text data. This paper merges an efficient online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means algorithm, using online update based on the well-known. Winner Take All competitive learning. It has been shown to be as efficient as SPKM, but much superior in clustering quality. The scalable clustering strategy was previously developed to deal with very large data bases that cannot fit into a limited memory and that are too expensive to read/scan multiple times. Using this method, one keeps only sufficient statistics for history data to retain (part of) the contribution of history data and to accommodate the limited memory. To make the proposed clustering algorithm adaptive to data streams, a forgetting factor is introduced here that applies exponential decay to the importance of history data. The older a set of text documents, the less weight they carry. The experimental results demonstrate the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams one needs to forget to be adaptive

Paper presented by yuanbinWu, xuanjing Huang [5] that define an opinion unit as a triple consisting of a product feature, an expression of opinion, and an emotional attitude(positive or negative).They use this definition as the basis for our opinion mining task. Since a product review may refer more than one product feature and express different opinions on each of them, the relation extraction is an important subtask of opinion mining. Introducing the concept of phrase dependency parsing segment an input sentence into “phrases” and links segment with directed arcs. The parsing focuses on the :phrases” and the relation between them, rather than on the single words inside each phrase. Because phrase dependency parsing naturally divides the dependencies into local and global, a novel tree kernel method has also been proposed.

III. TEXT MINING TECHNIQUES

More number of algorithms is available on text mining such as clustering, classification, information, extraction, summarization, visualization.

A. Information Extraction

Unstructured text is used for the information extraction. Information extraction software is identify the key, phrase, feature terms, word. It is very useful when deal with large volume of text. In traditional data mining techniques the information mined from the relational database. Information extraction is used to automatically extract the structured information from unstructured document. It is used more application such as structured search, opinion mining, sentiment extraction. In the information extraction model document converted to structured database which data mining techniques can be applied to extract the knowledge.

B. Information Retrieval

Information retrieval is retrieval of information from large amount of text based document. Information retrieval and data base handle different kinds of data. Some data base problem not present in this Information retrieval such as update, concurrency control, recovery, transaction management. Some common information retrieval problem not encountered in database system such as unstructured documents, appropriate search based keyword. Information access process is called information filter.

C. Text Summarization

Text summarization is compressed it’s input which specify human consumption. It contain individual document or group of

document. The output of the text summarization is human readable. It is decomposed in three main stage. These are Interpretation, Transformation, Generation. Interpretation of source text to obtain a text. Transformation of text representation into a summary representation. Generation of summary text from the summary representation.

D. Mining Structured Text

Mostly text on the internet contain explicit structural mark up and differ from the plain text. some markup indicate the document structure or format. This mark up is internal. some of external and give the hypertext link between the document. These information source give the additional benefit for mining web document. source of the information are extremely noisy. They involve arbitrary and unpredictable choice by individual page designers. Thus"web mining" is emerging as a new subfield, it is similar to the text mining but taking advantages of extra information available in the web document. Wrapper induction and document clustering and determining the authority of web document techniques are mostly used. Wrapper induction uses internal markup information increase the effectiveness of text mining in marked-up document. The remaining two, document clustering and determining the "authority" of web document, capitalize on the external markup information that is present in hypertext in the form of explicit links to other document.

E. Approaches To Text Mining

Using well-tested methods and understanding the results of text mining:- Once a data matrix has been computed from the input documents and words found in those documents, various well-known analytic techniques can be used for further processing which includes methods for clustering, factoring, or predictive data mining. Black-box approaches to text mining and extraction of concepts. There are text mining applications which use black-box methods to take out detailed meaning from documents with less human effort. These text-mining applications summarize large numbers of text documents automatically, retaining the core and most important meaning of those documents. Text mining as document search. The another approach of text mining is the automatic search of large numbers of documents based on key words or key phrases. This provides efficient access to Web pages with certain content. It searches very large document repositories based on varying criteria.

IV. TEXT MINING ALGORITHM

Various algorithms are available to effective classification and categorization in data mining.

A. K Nearest Neighbour

In the text mining domain the k nearest neighbour algorithm is a classical and frequently 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. 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. we consider each of the characteristics in our training set as a different dimension in some space, and take the value an observation has for this characteristic to be its coordinate in that dimension, so getting a set of points in space. We can then consider the similarity of two points to be the distance between them in this space under some appropriate metric. The way in which the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the k closest data points to the new observation, and to take the most common class among these. This is why it is called the k Nearest Neighbours algorithm.

B. Support Vector Machine

This approach is a one of most efficient 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. One-class SVM builds a profile of one class and when applied, flags cases that are somehow different from that profile. This allows for the detection of rare cases that are not necessarily related to each other. This is an anomaly detection algorithm which considers multiple attributes in various combinations to see what marks a record as anomalous. It first finds the "normal" and then identifies how unlike this each record is – there is no sample set. The algorithm can use unstructured data, text, also and use nested transactional data

C. Bayesian Classifier

This is 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. That is some time slow learning classifier additionally that do not produces the more accurate results. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

D. 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 efficient 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 efficient for cluster formation or categorization but comparative accuracy is much reliable than k-mean clustering. k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function.

V. CONCLUSION AND FUTURE WORK

In this paper various text mining techniques and algorithm are discussed for efficient and accurate text mining. Text mining is a techniques which is used to extract the high quality information or interesting information or knowledge from the text document which are usually unstructured form. Here in this work quite research field "Text mining" is discussed with it's various techniques which can be used such as summarization which can be produce the relevant information from the corpus. Classification is the supervised techniques which can be used to classify the new arrived document. Clustering is used ti divide the text into the clustering according to the similarity. It is a unsupervised learning which is no pre-defined input-output pattern are there. Extraction is basically used extract the structured information from unstructured text in which data mining techniques is used to extract the useful information from the document. In addition of that the efficient algorithms are also learned. In future the proposed technique is implemented using java technology and the comparative results are provided.

REFERENCES

[1] Shady shehata, Mohamed S. Kamel., "An efficient concept-based mining model for enhancing text clustering",IEEE transactions on knowledge and data engineering, vol. 22, no. 10, october 2010
[2] C. C. Aggarwal and P. S. Yu.,"A framework for clustering massive text and categorical data streams",‖ in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.
[3] M. Steinbach, G. Karypis, and V. Kumar., "A comparison of document clustering techniques",‖ in Proc. Text Mining Workshop KDD,2000, pp. 109–110.
[4] S. Zhong., "Efficient streaming text clustering",‖ Neural Netw.,vol. 18, no. 5–6, pp. 790–798, 2005.
[5] YuanbinWu, Qi Zhang, Xuanjing Huang, LideWu Fudan.,"phrase Dependency parsing for opinion mining" University school of computer science.
[6] S. Guha, R. Rastogi, and K. Shim., "rock: A robust clustering algorithm for categorical attributes",‖ Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
[7] 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
[8] D.Cutting, D. Karger, J. Pedersen, and J. Tukey., "A cluster-based approach to browsing large document collections",‖in Proc. ACM SIGIR Conf., New York, NY, USA, 1992,pp. 318–329.
[9] Vishal Gupta, Gurpreet S. Lehal, ―A Survey of Text Mining Techniques and Applications‖, Journal Of Emerging Technologies In Web Intelligence, VOL. 1, NO. 1, AUGUST 2009
[10] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, ―Effective Pattern Discovery for Text Mining‖, IEEE Transactions on Knowledge and Data Engineering. C Copyright 2010 IEEE

