International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue V, May- 2016

STUDY ON VARIOUS TECHNIQUES OF TEXT MINING

M.Arvindhan[1,*], *S.Prasanna*[2]
[1,*]saroarvindmster@gmail.com,2
[1,*] MAM School of Engineering,Trichy,Tamilnadu
[2]Sastra university -Srinivasa Ramanujan Centre, Kumbakonam, Tamilnadu
*Corresponding author

ABSTRACT

A new work and a challenging task in the emerging and enlarging field of research is text mining. Over a huge collection of text that are not arranged in proper order without any format and structure, an ability to achieve and extract certain kind of information and knowledge is termed as Text Mining and based on this work certain tools can be developed. Useful information from the raw and unformatted data can be gathered with the help of using tools that performs an analysis on large quantities of data. They can detect lexical and linguistic patterns found in those unformatted and structure free data. With the help of these obtained patterns, information is drawn up. The main aspect of text mining tools is that, intellectual answers will be offered and efficient text searches are performed with regard to the user. To expose the modernistic developments in this field and to review the basics of text mining is the motto behind this survey paper. In addition, different algorithms that belong to other areas like data mining and machine learning that are used for text mining are discussed.

Index Terms: Data Mining, Text mining, Machine learning, Clustering, Classification.

1. INTRODUCTION

Assembling of useful details and information and removing unwanted data from larger sets of data is referred as Data Mining. Apart from collection of information, experts found it is possible that large amount of data can be stored at low cost. To expel out unwanted data and to extract relevant information from large amount of data is the main target of this generic task that is termed as Data Mining. Knowledge discovery from text (KDT), which deals with the machine supported analysis of text, is another term that is used to represent Text Mining. To extract explicit, implicit concepts and semantic relations between the concepts is the main role of KDT. Only humans have the capability to overcome the obstacles such as correct spell, slang and correct meaning that computers cannot easily handle and distinguish between patterns in linguistic nature. But to operate large volumes of text at high speed, we need a computerized system. Text mining is used for managing the knowledge and Human resource,

Customer Relationship, Technology Watch, Natural Language Processing and Multilingual Aspect.

This paper, describes data mining and text mining as a whole process in second section. Third section explains about the different application of text mining, clustering and classification and also about text mining tools. In Last section, explains machine learning in text mining.

2. TEXT MINING AND DATA MINING

Mining the text we look for patterns in text whereas we look for patterns in the data while mining the data. However, the similarity between the two conceals exact difference. Data mining can be added fully characterize as the removal of implicit, before unidentified, and potentially helpful information from data [Witten and Frank, 2000]. The following information is implied in the data as an input: it is concealed, unidentified, and could scarcely be extracted devoid of option to the automated techniques of mining the data. However, the data to be extracted is clearly and explicitly declared is down in Text mining. It's not concealed at all the majority authors go to enormous pains to make sure that they express themselves clearly and unmistakably and, from a human point of sight, the human resource restrictions from the intellect in which the correct knowledge from the text is "not identified in early

stages". Here the actual problem is information is not couched in a manner that is amenable for automation. Though there is a clear difference philosophically, since the computer's point of vision the troubles are fairly related. Text is presently as opaque as rare data once it comes to extracting information-almost certainly more so. One more obligation that is familiar to both mining the data and text is that the information extracted should be "potentially useful". It gives a meaning "actionable" capable of providing a basis for actions are to be automated. In the case of data mining, this notation can be expressed in a relatively domain-independent way: actionable patterns are ones such that on a new data from a similar source non-trivial predictions can be made. Performance preserve be precise through together with successes and failure, statistical technique can be useful to compare different data mining methods on the identical problem, and so on. However, in many text mining situations it is far harder to characterize what the "actionable" means in a way that is independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success.

In most of the applications based on data mining, the term "potentially useful", imparted in a different interpretation: the key for success is that the information extracted must be useful in that it helps to explain the data. It is necessary whenever the result is used for consumption purpose by human beings rather than (or as well as) a basis for automatic action. This criterion is less applicable to text mining since, dissimilar data mining, the input itself is understandable. Text mining with understandable output is tantamount to summarizing salient features from a huge body of text, which is a subfield in it that possesses right text summarization.

2.1 TEXTMINING PROCESS:

Data mining is the process of finding and segregating, through automatic or semiautomatic resources, of huge quantity of data in order to determine meaningful pattern and rules. Data mining is an interdisciplinary sub-field of computer science that involves computational process of large data sets patterns discovery. The goal of this advanced analysis is to takeout information from a set of data and transforms them toan understandable structure for further use. The methods used are Text mining, machine learning, statistics, and database systems. Data mining is also stated as essential tasks that comprise intelligent methods in order to extract the data patterns. The steps involved in the overall process of the text mining are depicted in the Fig.1.

A. TEXT PROCESING:

There are number of sub steps in the text processing are as follows:

1) Tokenization: Text document has a collection of sentences. This step divides whole statement into words by removing spaces, commas etc.

2) Stop word Removal: This step involves removing of HTML, XML tags from web pages. Then process of removal of Stop words like "a", "of" etc. is performed. Finally word stemming is applied.

3) Stemming: The root/stem of a word is determined by these techniques. It converts words to their roots/stemse.g. Flying, Flew word to Fly.

B. TEXT TRANSFORMATION / FEATURE GENERATION

Text transformation of Text is done by converting the document into bag of words or vector space document model notation. It can be used further for various tasks for effective analysis.

C. FEATURE ELECTION / ATTRIBUTE SELECTION

This phase mainly used for removing features which are considered irrelevant for mining purpose. This advantage of this procedure is to produce a smaller dataset size, less computations and minimum search space required.

D. TEXT MINING METHODS

There are different text mining methods as in Data mining had been proposed such as: Clustering, Classification, Information retrieval, Topic discovery, Summarization, Topic extraction.

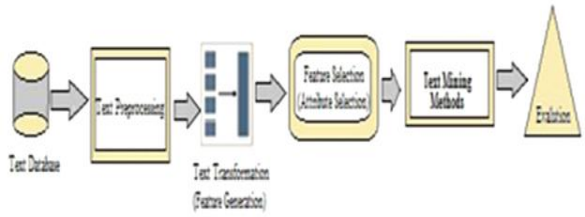

Fig.1. Text mining process flow

D. Interpretation or Evaluation

This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

3. APPLICATION OF TEXT MINING

Text mining has a high worth in commercial. For analyzing large amount of unstructured documents for the purpose of extracting interesting and non -trivial pattern or knowledge it's a rising application.

Text mining has much domain specific applications.Here some of the applications are explained:

1) Customer Profile Analysis: Text mining is basically used to collect the instance and occurrence of key terms in blocks of data such as articles, Web pages. Important data drilling tools such as topic structures and semantic network are converted by software from unstructured data formats. The general tone of complaints, reason for complaining can be learned through the semantic network. Through semantic weight finding the common words used and their relationship to other words in text can be found easily.

2) Security applications: For security applications there are many test mining software's, especially monitoring and analysis of online text sources for national security purpose. It also involved in the study of encryption/decryption of text.

3) Biomedical Application: In biomedical and it's intended for identification and classification of technical terms in the domain of molecular biology's corresponding to concepts text mining are used.

4) Company Resource Planning: Mining Company's information and correspondences for activities, so its resource status and problems reported that can be handled properly and future action planned can be design.

5) Competitive intelligence: Enabling the companies to organize and modify the company strategies according to present market demands and the opportunities based on the information collected by the company about themselves, the market and their competitors, and to manage enormous amount of data for analyzing to make plan.

6)Customer Relationship Management (CRM): Rerouting exact needs automatically to the appropriate service or providing immediate answers to the most frequently asked questions.

7) Technology watch: Identification of the applicable Science and Technology literatures and extraction of the required Information from these literatures, text mining techniques are used extensively.

8) Organize repositories of document-related meta-information: Automatic text classification methods are used to create structured metadata, which is used for searching and retrieving relevant documents based on query.

9) Human Resource Management: Mainly with applications aiming at analyze staff's opinions and monitoring satisfaction of employees as well as reading and storing of CVs for the selection of new personnel TM is used.

3.1Classification&Clustering:

Classification is a part of data mining technique used to predict group membership for data instances. For example, if you use classification to predict the weather on a particular day will be "sunny", "rainy" or "cloudy". K nearest neighbor algorithm (KNN), Bayes algorithm, Support Vector Machine algorithm (SVM), decision tree algorithms, neural network algorithm (Nnet) are some of popular classification techniques. The traditional KNN text classification algorithm is used for all training samples for classification. Clustering is a process of partitioning a set of data (or objects) into a set

of meaningful sub-classes. Popular clustering technique includes k-means clustering and expectation maximization (EM) clustering. Clustering is mostly confused with classification, but there are some differences between them. In distributing, the objects are subjected to pre-defined classes, where as in clustering the classes are also to be explained and defined. Hence, in Data Clustering, the information that is logically similar is physically stored together. Objects that contain same properties are placed in a class of objects in clustering and a makes the entire class is made available in the disk just because of single access.

3.2 Text Mining tools:

Text mining is concerned with discovery of structure and patterns in unstructured data – typically text. There are a lot of dissimilar approach toward this task, a quantity of focus on ancillary structure such as taxonomy and ontologies, a quantity of focus on semantics and NLP, while the remaining part use various different algorithms toward categories and summarize.

Among well-known open source data mining tools offering text mining functionality is the **Weka** (Witten and Frank 2005) suite, a collection of ML(machine learning) algorithms for data mining tasks also offering classification and clustering techniques with extension projects for text mining, like **KEA** (Witten et al. 2005) for keyword origin. It provides good API support and has a broad user base. Then there is **GATE** (Cunningham et al. 2002), Text Mining Infrastructures in **R** an established text mining framework with architecture for language processing, informationextraction, ontology management and machine learning algorithms. It is fully written in Java. Other tools are **Rapid Miner**, a system for knowledge discovery and data mining, and Pimiento (Adeva and Calvo 2006), a fundamental Java structure for text mining. However, a lot of existing open-source products be likely to offer rather Specialized solutions in the text mining context, such as Shogun (Sonnenburg et al. 2006),a toolbox for string kernel, or the Bow toolkit (McCallum 1996), a C documents useful for statistical text analysis, language modeling and information retrieval. In R the extension package **ttda** (Mueller 2006) provides some methods for textual data analysis.

4. MACHINE LEARNING FROM TEXT ANALYSIS

Basically machine language(ML) deals with construction of computer programs that automatically that improve with experience and it comes as supervised and unsupervised.In first computer programs capture structural information and derive conclusion from previously labeled examples (instance,points).in latter partit finds groups in data without relying on labels.
Techniques in ML can basically divide into fourdistinctareas:classification,clustering,associati on learning and numeric prediction. In text categorization(tc) classification is applied to text in the task of automatically sorting a set of documents into categories from a predefined set. Documentclassification is employed in text filtering, web page categorization, sentiment analysis, etc. further used in parts of text depending on concrete application, e.g. document segmentation or topic tracking. In ML, before being applied to sorting unseen texts, they are trained on previously sorted data. naivebayes, k-nearest neighbor, and support machines are some popular classifiers.

The generalization of evidence produced by dataset is used to find models in classification concern. In clustering models are discovered by finding group of data points those satisfy objective criterion. e.g during minimizing similarity of points from different clusters.

5. CONCLUSION

Text mining is an interesting field, which gives scope to the researchers to explore the real world textual corpus with different perspective and to extract thought provoking text patterns and distribution with global statistical data. The text mining tools and applications makes the scientific community to pay attention towards text engineering to bring out fascinating results from corpora.

References

1. Adeva, J.J.G. and Calvo, R.A., 2006 Mining Text with Pimiento. IEEE Internet Computing.10(4):27- 35.

2. Sonnenburg,S., Raetsch G., Schaefer. C.,Fraunhofer, DE., 2006. Journal of Machine Learning Research 7:1531–1565.
3. McCallum A. and Culotta A.,Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text.
4. Müller E., Günnemann S., Färber I., SeidlT.: 2012. Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data Tutorial at 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD 2012**), Kuala Lumpur, Malaysia.
5. Witten, I. H. (Ian H.) Data mining: practical machine learning tools and techniques. 3rd ed. / Ian H. Witten, Frank Eibe, Mark A. Hall. p. cm. (The Morgan Kaufmann series in data management systems) ISBN 978-0-12-374856-0 (pbk.) 1. Data mining. I. Hall, Mark A. II. Title. QA76.9.D343W58 2011 006.3′12—dc22 2010039827
6. Witten, I. H. (Ian H.) Data mining : practical machine learning tools and techniques / Ian H. Witten, Eibe Frank. – 2nd ed. p. cm. – (Morgan Kaufmann series in data management systems) Includes bibliographical references and index. ISBN: 0-12-088407-0 1. Data mining. I. Frank, Eibe. II. Title. III. Series. QA76.9.D343W58 2005 006.3–dc22 2005043385
7. Althoff, K.D., Bergmann, R Minor, M., Hanft. A.Advances in Case-Based Reasoning: 9th European Conference, ECCBR 2008, Trier, Germany, September 1-4, 2008, Proceedings 2008

8.Xue Li, Osmar R. Zaiane, Zhanhuai LiAdvanced Data Mining and Applications: Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006, Proceedings.