International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

A Study on Recent Issues in Text Pre-processing and Classification Techniques

Dr. Ramalingam Sugumar

Professor & Deputy Director, Christhu Raj College, Tamil Nadu, India

ABSTRACT

Data Mining is a significant research area in the field of computer science. Data mining techniques are applied to textual data sets is known as Textual Data Mining (TDM). The TDM consists of two stages Pre-processing and post-processing. In TDM pre-processing faces several issues in various stages such as Tokenization, Stop Word Removal and Stemming. The stemming is one of the pre-processing technique in text mining. It is mainly used to removing inflectional and derivational endings in order to reduce word forms to a common stem. The stemming involves text processing task includes information retrieval, text mining, and natural language processing. In this study, discuss recent issues in text pre-processing classification algorithms in text mining. The study is to show the merits and demerits of text mining techniques.

Keywords : *Text Data Mining, Text Mining, Stemming algorithms, classification algorithms Truncating, Statistical.*

I. INTRODUCTION

Data mining is a process of discovering hidden patterns and information from the existing data. Data mining requires different algorithm for manipulating and analyzing data from large data repository. It applied techniques for successfully detecting fraud or lies text based communication in on-line[1].There are numerous data supported in data mining such as sequence data, sequential data, time series, temporal, spatio-temporal, audio signal, video signal etc.[2]. Among all these types of data, particularly data mining supports text data for representing the document.

Due to vast usage of text data, text mining has become a challenging research area in data mining. Hence, It organizes and classifies the text document which is considered as a challenging issue. Classification of the text documents or information is stated as text classification. In text classification, several steps are involved such as text preprocessing, feature extraction, feature selection and classification methods. Text classification methods are successfully applied to different areas such as topic detection, spam e-mail filtering, clinical medicine, sentimental analysis and web page classification Most commonly used pre-processing steps are tokenization, stop word removal, lowercase conversion and stemming.

II. Related Work

Text Mining

An effective text mining is predicted with sophisticated preprocessing techniques. Text preprocessing is one of most important steps in Text mining. Text preprocessing is necessary to filter the information using relevant condition depends upon the text. Preprocessing consists of several stages such as data cleaning, data integration, description and summarization, attribute transformation and data reduction. Natural Language Processing is one of task in frequently used in text preprocessing. After, preprocessing the information posted to classification methods. In text mining, text preprocessing usually follow the steps such as tokenization, stop word removal, lower case conversion and stemming.

Tokenization is the task to segment the sentence into words, phrases or other meaningful parts which are stated as tokens. Stop word is a process for removing irrelevant words such as articles, preposition, conjunction and so on. Stemming is used to find the root of derived word in a document. Therefore, preprocessing task is commonly used to extract the features and feature selection.

The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific (pre) processing methods and algorithms are required in order to extract useful patterns. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. Text mining is the process of discovering information in text documents. Pattern mining involved in text mining for discovering patterns from large collection of text database. SP. Ruba et al. [3] proposed APOST for increasing the performance of stemming. The performance of APOST stemmer examined with sample vocabulary downloaded from the website http://snowball.tartarus.org/algorithms/english /voc.txt. It contains distinct words, arranged into "conflation groups". Some of them are incorrect words. Pattern mining techniques can be used to find various text patterns such as sequential patterns, frequent item sets and so on. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined deals with together into ingle workflow. We will now describe in more detail each of these areas and how, together, they form a text-mining pipeline.

Information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents [4]. Natural language processing (NLP) the automatic processing and analysis of unstructured textual information. One direction of NLP research relies on statistical techniques, typically involving the processing of words found in texts [5]. In general, a document is broken up into chunks (e.g., sentences or paragraphs), and rules or patterns applied to identify entities. Extracting the information by using different type of extraction [6].

Concept of Stemming

Stemming process involves affix removal algorithm which removes prefixes and suffixes of the word in the document[7,8].Stemming can be error in two ways)Over stemming which removed too much of words ii)Under stemming which removed little much of words. The stemming algorithm can be divided in to two groups: Truncating and Statistical [9].There are numerous stemming algorithms are used to trimming the words, notably the Lovins stemmer, Porter stemmer , Paice/Husk and Dawson stemmer. All these algorithms used under the truncating method. The statistical method supports YASS, N-Gram and HMM stemmer. Broadly, stemming algorithms can be classified in to truncating methods and statistical methods.

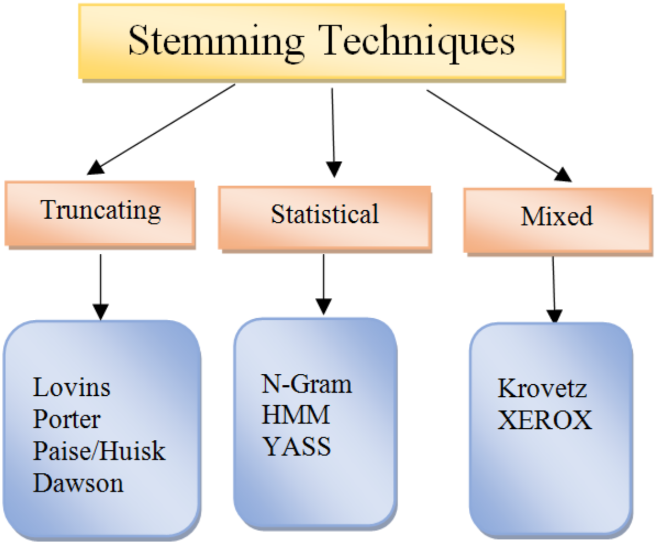

Figure 1 : Classification of stemming techniques.

Differenrt StemmingAlgorithms

Truncating Method

Porter's algorithm[10] is not applicable for removing a suffix when the stem is too short [14].It calculates the words based on the Vowel (A, E, I, O, U) Consonant (Other than Vowel) Pairs, and it is denoted by 'm'. By using multiple step Process it successively removes the short suffixes, instead of removing a single longest possible suffix.

Porter stemmer cannot handle irregular verbs. It has at least five steps and sixty rules for generating stem. It consumes more time to execution. It is suitable for American English but we follow the British English. It cause Over-stemming Problem.

The basic idea of Lovins stemmer is to remove suffix from the word. This algorithm involved list of 294 suffices, 29 conditions and 35 transformation rules which has been used for longest match principal. The word is recoded using different table after the ending is removed. This adjustment makes convert these stems into valid words.

The benefits of this algorithm is it is very fast and can handle irregular plural words like mouse and mice also removal of double letter words like 'running' being transformed to 'run'.

The drawback of this algorithm are time consuming, all suffixes are not available, and its size is bigger than porter by involved number of transformations based on the letters within the stem.

Paice/Husk Stemmer [11] is an iterative stemmer which removes or replaces the ending from a word in an infinite number of steps. It maintains a table of rules. When the word is processed, this algorithm uses the index for applying first rule to the last letter of word. If the rule is accepted, the resultant is applied to the word otherwise the next rule index incremented by one and applied the next rule. Hence, it is over stemming will occur.

Dawson Stemmer is an extension of the Lovins stemmer which attempt to refines the rules and techniques of Lovins stemmer. It has a list of 1200 suffixes and also a single pass context-sensitive suffix removal stemmer. It corrects the basic error which has done by Lovins stemmer.

The advantages of Dawson stemmer is fast in execution and covers more suffixes than Lovins.
The benefits are very complex and standard implementation is poor.

Statistical Method

It is popular and effective approach in information retrieval. Some recent studies [12] show that statistical stemmers are good alternatives to rule-based stemmers. In this method word Stemming is done after applying certain Statistical Techniques like N-Gram, HMM, YASS. This type of stemmers is based on statistical Analysis and techniques.

N-Gram clustered the related pairs of words. This method based on digram or trigram which represents pair of consecutive letters. So it is called as [13,14]N-Gram method. This method measures association between the pairs of terms based on shared unique digrams. For calculating this association measures use Dice's coefficient.

Once the unique digrams for the word pair have been identified and counted, a similarity measure based on them is computed. The similarity measure used is Dice's coefficient, which is defined as:

$$S = \frac{2C}{A + B}$$

where A is the number of unique digrams in the first word, B the number of unique digrams in the second, and C the number of unique digrams shared by A and B.

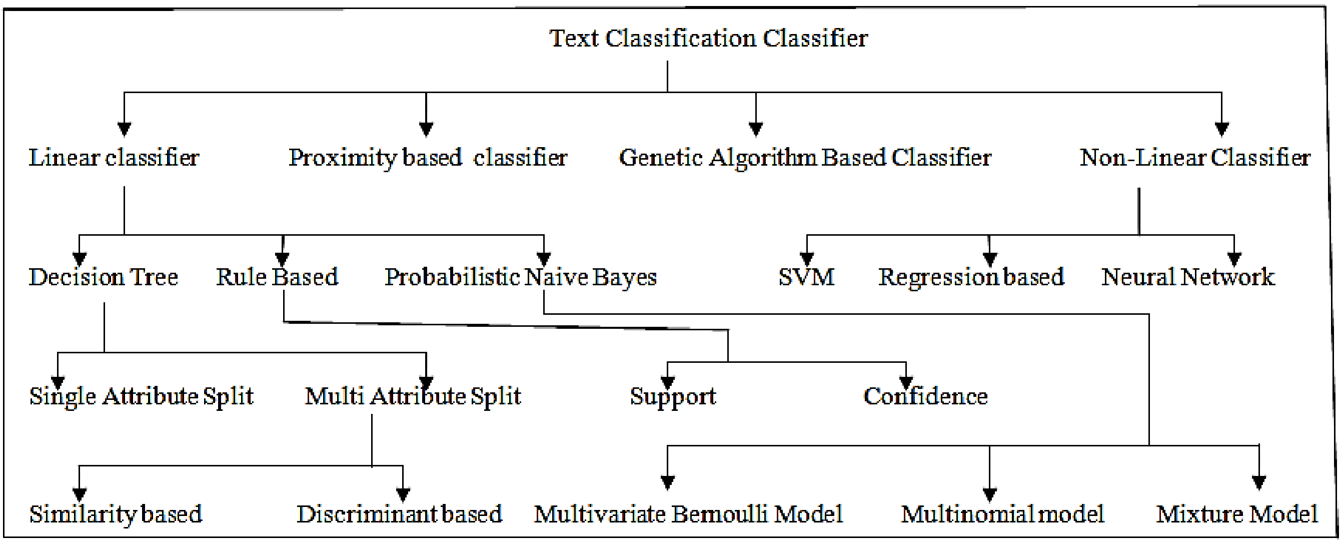

Figure 2 : Text lassification algorithms

where A is the number of unique digrams in the first word, B the number of unique digrams in the second, and C the number of unique digrams shared by A and B. YASS[15,16] is an acronym for Yet Another Suffix Striper. This approach was proposed by. HMM method does not require a prior linguistic knowledge of the dataset with based on unsupervised learning. The benefits of this method are it is unsupervised and hence knowledge of the language is not required. The disadvantage is complex method for Implementation. Over stemming may possible. N-Gram Stemmer not a very practical method. So, no time efficient. It requires significant amount of space for creating and indexing N-grams. YASS stemmer is difficult to decide a threshold for creating clusters. It requires significant Computing power.

D. Different Classification Algorithms

Decision Tree: Decision trees [17] represent one of the main techniques for discriminant analysis in data mining and knowledge discovery. They predict the class membership of an instance using its measurements of predictor variables. The most popular algorithms for decision tree induction are based on top-down greedy search. First, the test attribute is decided for the root node. Instances are split through the tree from the root node to a leaf node, which provides classification of a given instance. A teach non-terminal node through which the instance passes, one (or many) attribute of the instance is tested and the instance is moved down to the branch that corresponds to an outcome of the test. The process is recursively repeated for each branch. When to stop partitioning and create a leaf node is still one of the major problems in the area. Classification trees have many advantages that make them applicable in various scenarios, particularly when the data does not satisfy the rigorous assumptions required by more traditional methods. In this paper, the following facts are significant.

Support Vector Machine: The aim is to determine those support vectors that maximize the distance or ‘margin’ between the boundaries and their resultant prototype vector, hypothetically producing the greatest separation of the opposing training samples and increasing classification accuracy. Extension of the basic concept to include any number of features is also possible, in which case the decision surface takes the form of a multidimensional hyper plane. It is also not restricted to just linearly separable data and requires no manual input, with all parameters capable of being automatically tuned, though the actual quantity and type of variables depend on the specific components employed. The non-linear mapping induced by the feature functions is computed with special non-linear functions called kernels. In this case, the solution is defined as a weighted sum of the values of certain kernel function evaluated at the support vectors. Support Vector Machine (SVM) has major drawbacks of increased speed, memory space and computational cost.

Neural Network: In neural networks, Extreme learning machine (ELM) has gained increasing interest from various research fields recently. Apart from classification and regression, ELM has recently been extended for clustering, feature selection, representational learning and many other learning tasks. These newly emerging algorithms greatly expand the applications of ELM. From implementation aspect, hardware implementation and parallel computation techniques have substantially speed up the training of ELM, making it feasible for big data processing and real-time reasoning. Due to its remarkable efficiency, simplicity, and impressive generalization performance, ELM have been applied in a variety of domains, such as biomedical engineering, computer vision, system identification, and control and robotics. Conclusion In this study, detailed discuss about text pre-processing and text classification techniques. This study is focus four affix removal stemming algorithm, three statistical stemming algorithm and classification techniques. Finally, Porter stemmer is faces several issues such as cannot handle irregular verbs, over-stemming and under-stemming. Support Vector Machine has major drawback such speed, time consumption and computational cost. In future, to rectify identified problems to better classification result.

III. REFERENCES

[1]. Ning Zhong,Yuefeng Li and Sheng-Tang Wu, Effective pattern Discovery for Text Mining.,IEEE Transactions on knowledge and data engineering .,vol 24,jan 2012.

[2]. M.S.B. PhridviRaj, C.V. GuruRao., Data mining – past, present and future – a typical survey on data streams, Elsevier 2013.

[3]. S.P. Ruba Rani, B.Ramesh and Dr.J.G.R.Sathiaseelan, "An Increasing Efficiency of Pre-processing using APOST Stemmer Algorithm for Information Retrieval", Journal of Emerging Technologies and Innovative Research, Volume 2, Issue 7, pp.3219-3223, July 2015.

[4]. ChristieM.Fuller,David P.Biros,Dursun Delen., An investigation of data and text mining methods for real world deception detection,Elsevier 2011.

[5]. R. Sagayam, S.Srinivasan, S. Roshni., A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques., International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5,September 2012.

[6]. Manning, C. and Schutze, H. Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[7]. Scaling Information Extraction to Large Document Collections, Eugene Agichtein, Microsoft Research, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

[8]. Anjali Ganesh Jivani, A Comparative Study of Stemming Algorithms, IJCTA | NOV-DEC 2011.

[9]. http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap08.htm ,CHAPTER 8: STEMMING ALGORITHMS , W. B. Frakes, Software Engineering Guild, Sterling, VA 22170.

[10]. Willett, P. (2006) The Porter stemming algorithm: then and now. Program:Electronic library and information systems, 40 (3). pp. 219-223.

[11]. Paice, C., Husk, G., Another Stemmer, ACM SIGIR Forum 24(3): 566, 1990.

[12]. S.Santhana Megala, Dr.A.Kavitha Dr. A.Marimuthu, Improvised Stemming Algorithm – TWIG, 2013, IJARCSSE.

[13]. J. B. Lovins, "Development of a stemming algorithm," Mechanical Translation and Computer Linguistic., vol.11, no.1/2, pp. 22-31, 1968.

[14]. K.K. Agbele, A.O. Adesina, N.A. Azeez , & A.P. Abidoye, ContextAware Stemming Algorithm for Semantically Related Root Words., AfricanJournal of Computing & ICT June, 2012 .,IEEE.

[15]. M. Porter (1980). An Algorithm for Suffix Stripping.

[16]. Program, vol. 14, no. 3, pp: 130 – 137.

[17]. Gobinda Kole, Pabitra Mitra andKalyankumar Datta. "YASS: Yet another suffix stripper". ACM Transactions on Information Systems.Volume 25, Issue 4. 2007, Article No. 18.

[18]. Yang Shao and Ross S. Lunetta, "Comparison of support vector machine, neural network, CART algorithms for the land-cover classification using limited training data points", ISPRS Journal of Photogrammetry and Remote sensing, Elsevier, Volume 70, pp.78-87, 2012.