Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.165 – 171

RESEARCH ARTICLE

Evaluation of Stemming Techniques for Text Classification

S.P.Ruba Rani[1], B.Ramesh[2], M.Anusha[3], Dr. J.G.R.Sathiaseelan[4]

[1,2,3,4]Department of Computer Science, Bishop Heber College, Tiruchirapalli, TN, India

[1] sp.rubarani@yahoo.com; [2] ram.73110@gmail.com; [3] anusha260505@gmail.com; [4] jgrsathiseelan@gmail.com

Abstract— Text mining is perceived as a process of extraction of meaningful information from textual document repositories. The stemming technique is one of the preprocessing techniques in text mining, which truncate inflectional and derivational endings, it reduces word to fetch common stem. It involves text processing task that includes text mining, information extraction and natural language processing. Usually, most of the document contains morphological variants, hence stemming involves removal of unwanted affixes from the document. In this paper, various stemming techniques are analysed and classified hence study shows the benefits and limitations of recent stemming techniques.

Keywords— Text Mining, Stemming techniques, Decision based method, Statistical method.

I. Introduction

Data mining is a process of discovering hidden patterns and information from the existing data. Data mining [1] requires different algorithm for manipulating and analysing data from database. They are several techniques are available in data mining for analysing data such as clustering, classification, decision tree, neural network and genetic algorithm. There are numerous data supported in data mining such as sequence data, sequential data, time series, temporal, spatio -temporal, audio signal, video signal etc. Among all these types of data [2], particularly data mining supports text data for representing the document. A document consists of collection of words which includes stop words (is, the, are etc).Many words used in the text are morphological variants which based from the root form e.g. connection /connect, combining /combine, preferences /preferred/prefer. Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management. Text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results.

Text classification [3] is one of the major parts in text mining. Nowadays, handling textual documents is a great challenge. Text classification uses several key classification algorithms, e.g., decision trees, pattern (rule)-based classifiers, support vector machines, naïve Bayesian classifier and artificial neural networks. The feature extraction method is based on the probabilistic topic model using graph based classifier for text classification [4]. Text classification applied in many fields. Biological genetic algorithm for instance selection of text classification in medical field [5].

Preprocessing is one of the methods for text classification. Preprocessing task involves four common steps such as tokenization, stop-word removal and lowercase conversion and stemming. These steps implemented in different manner according to the document representation. In text preprocessing task, tokenization is the process of finding words separating them into words, phrases or other meaningful parts known as tokenization. Filtering of important and relevant words from the list of words which were the output of tokenization is known as stop word removal e.g. removes the stop words like ‘have, it, can, but, to, also from the document. Uppercase to lowercase conversion is important for text classification. Stemming removes the prefixes and suffixes of each word which reduces words variants to its root form.

II. LITERATURE REVIEW

A. Text Mining

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 preprocessing 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 .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 a single workflow. We will now describe in more detail each of these areas and how, together, they form a text-mining pipeline.

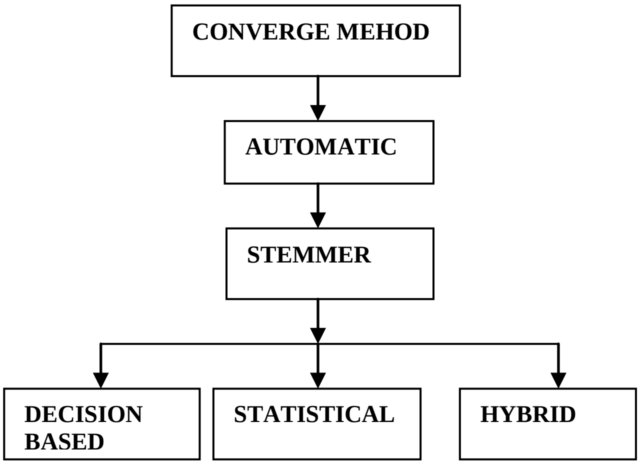
Information retrieval [6] is concerned with the organization and retrieval of information from a large number of text-based documents. Natural language processing (NLP) [7] the automatic processing and analysis of unstructured textual information. One direction of NLP research relies on statistical techniques, typically involving the processing of words found in texts. In general, a document is broken up into chunks (e.g., sentences or paragraphs), and rules or patterns applied to identify entities.

B. Concept of Stemming technique

Stemming is the process of reducing the words in to root form effectively. For example, the user enters the term ‘examination’ in IR query, the system retrieve ‘examined’ ‘exam’ as the resultant word. Stemming Process involves affix removal algorithm which removes prefixes and suffixes of the word in the document. Stemming can be error in two ways, i) Over stemming which removed too much of words ii) Under stemming which removed little much of words. Stemming and stop word removal method can be avoid Named entity problem[16].The stemming algorithm can be divided in to three groups such as Decision based, Statistical and Hybrid .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 decision based method. The statistical method supports YASS, N-Gram and HMM stemmer.

Porter stemmer is the most common algorithm for English stemming. Improved version of the original porter stemming algorithm for English language is proposed in [8].The performance of this stemmer can be calculating the number of under stemming and over stemming errors. There are various available stemming alternatives widely used to enhance the effectiveness and efficiency of information retrieval [9].

Conflation technique is needed for getting a single term from different morphological variants. Automatic conflation is known as stemming. Automatic conflation can be categorized into Affix removal, Successor variety, Table Lookup and n-gram [10]. The measuring of strength and accuracy of stemming algorithm is needed for best way to retrieving the text[11].The stemming method can be applied with automated query generation components(RDBMS) for information extraction[15].



C. Different Stemming Algorithms

Decision Based Method

It relates to removing the suffixes or prefixes of a word. It truncated a word at nth position keeping the n letters and removes the rest based on some rules and conditions. This type of stemmer is one of the most popular techniques.

1. Porter Stemmer

The main idea of porter stemmer uses suffix stripping in English Language. This stemmer consists of five or six steps depend upon the process that is used to give the final stem. Original algorithm consists of only five steps. Each step applied the rules, and conditions also involved. If the rule is correctly accepted, the suffixes are automatically removed according to the condition, and the next step performed. The rules and conditions end at the resultant stem [12].
The porter steps are:

1) <Suffix> → <new suffix>

SSES → SS Crosses → Cross

IES → I Studies → Studi

SS → SS Caress → Caress

S → S Dogs → Dog

2) <Condition> <suffix> → <new suffix>

If rule (m>0) EED → EE

Agreed → Agree

Porter’s algorithm is not applicable for removing a suffix when the stem is too short [16].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. The benefit of porter stemmer is to apply minimum number of steps than other stemmer, less error rate, lighter than lovins stemmer. The limitation of this method is it follows at least five steps, 60 rules and leads to time consuming.

2. Lovins Stemmer

Lovins stemmer [13] is to remove suffix from the word. This algorithm involved list of 294 suffices 29 conditions and 35 transformation rules which have 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 limitations 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.

3. Paice/Husk Stemmer

Paice/Husk 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. The benefit of this algorithm is very simple than other stemmer. Iteration takes care of deletion and replacement. The limitation is over stemming will occur.

4. Dawson Stemmer

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

Statistical Method

It is popular and effective approach in information retrieval. Some recent studies show that statistical stemmers are good alternatives to decision-based stemmer. 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.

1. N-Gram

N-Gram clustered the related pairs of words. This method based on digrams or trigrams which represents pair of consecutive letters. So it is called as 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.*

TABLE I
SIMILARITY MEASURE OF DIGRAMS AND TRIGRAMS FOR WORDS

Steps	**Example Word : Constructor and Destructor**		
	Word Calculation	**Di-Grams**	**Tri-Grams**
1	Unique N-gram of Word 1	*C CO ON NS ST TR RU UC CT TO OR R*	**C *CO CON ONS NST STR TRU RUC UCT CTO TOR OR* R**
2	Unique N-gram of Word 2	*D DE ES ET TR RU UC CT TO OR R*	**D *DE DES EST STR TRU RUC UCT CTO TOR OR* R**
3	A = Unique N-gram of Word 1	12	13
4	B = Unique N-gram of Word 2	11	12

Steps	**Example Word : Constructor and Destructor**		
	Word Calculation	**Di-Grams**	**Tri-Grams**
5	C = Shared Unique Words	7	8
6	Dice Coefficient= (2C)/(A+B)	0.60869	0.64

Note: '*' denotes padding space

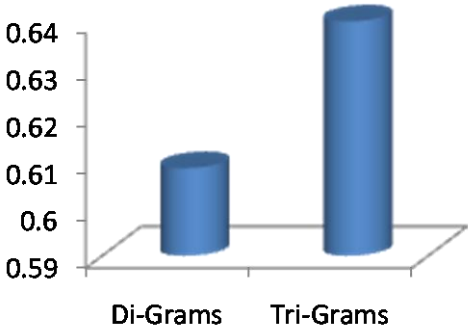


Fig 1. Similarity measurement of Di-Grams and Tri-Grams

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*. For the example above, Dice's coefficient would equal (2 x 7) / (12 + 11) = 0.60869. Such similarity measures are determined for all pairs of terms in the database. Once such similarity is computed for all the word pairs they are clustered as groups. The value of Dice coefficient gives us the hint that the stem for these pair of words lies in the first unique 8 digrams.The similarity measurement is shown in fig 1.

2. YASS

The acronym of Yass is Yet Another Suffix Striper. This approach was proposed by [14].The yass stemmer generated by clustering a lexicon. It performs without any linguistic input is comparable to that obtained using decision-based stemmers such as Porter's. This stemmer comes under the category of statistical as well as corpus based. It does not rely on linguistic expertise.It is based on clustering-based approach to discover equivalence classes of root words and their morphological variants. A set of string distance measures are defined, and the lexicon for a given text collection is clustered using the distance measures to identify these equivalence classes.

The benefits of this method is used for any language without knowing its morphology and based on hierarchical clustering approach and distance measures. The disadvantage is requiring more computing power.

3. HMM

HMM represents finite-state automata based on the concept of the Hidden Markov Model (HMMs) where transitions between states are ruled by probability functions. A hidden Markov model is an extension of a Markov process where the observation is a probabilistic function of a state.The new state produces a symbol with a given probability in each transition. This model was proposed by Melucci and Orio.

This method does not require a prior linguistic knowledge of the dataset with based on unsupervised learning. In automata graph, Viterbi coding is used to find the most probable path in order to probability of each path can be computed.

In HMM stemmer, word can be considered the result of a concatenation of two sub sequences: a prefix and a suffix. The states are divided in two disjoint sets: initial can be the stems only and the later can be the stems or suffixes. Transitions between states define word building process. There are some assumptions that can be made in this method:

1. Initial states always start with a stem.

2. Transitions from suffix state to stem state always have a null probability - a word can be only a concatenation of a stem and a suffix.

3. Final states belong to both states - a stem can have a number of different derivations, but it may also have no suffix.

The benefit of this method is it is unsupervised without knowledge of the language. The disadvantage is little complex.

D. Comparison between the algorithms

TABLE I
TRUNCATING (AFFIX REMOVAL) METHOD

Truncating method	
Algorithms	**Limitations**
Porter Stemmer	1)The stems produced are not always real words. 2) It has at least five steps and sixty rules for generating stem. 3) Time consuming. 4) It is suitable for American English but we follow the British English. 5)It causes Over-stemming Problem.
Lovins Stemmer	1) Time consuming. 2) All suffixes are not available. 3) Unreliable and fails to form words from the stems frequently.
Paice / Husk Stemmer	1) It is heavy algorithm. 2) Over stemming may possible.
Dawson Stemmer	1) Very complex. 2) Lacks a standard Implementation.

TABLE II
STATISTICAL METHOD

Statistical method	
Algorithms	**Limitations**
HMM Stemmer	1) A complex method for Implementation. 2) Over stemming may possible.
N-Gram Stemmer	1) Not a very practical method so no time efficient. 2) Requires significant amount of space for creating and indexing N-grams.
YASS Stemmer	1) Difficult to decide a threshold for creating clusters. 2) Requires significant Computing power.

III.CONCLUSION

This paper suggests four affix removal stemming algorithms and three statistical stemming algorithms. This study presented comparison of various stemming methods. Stemming technique significantly increases the retrieval results for both decision based and statistical approach. Decision based approach is mainly used for some specific language where as statistical based approach is language independent. The main difference between them is decision based accomplished by using set of rules and statistical done by using set of calculations. Both techniques are not fully produce 100% output, but are good enough to use for text mining.

REFERENCES

[1] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE, 2014.

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

[3] Francesco Colace, Massimo De Santo, Luca Greco and Paolo Napoletano, "Text classification using a few labeled examples", Computer in Human Behavior 30(2014)689-697, Elsevier, 2013.

[4] Alper Kursat Uysal and Serkan Gunal, "The impact of preprocessing on text classification", Information Processing and Management 50(2014) 104-112, Elsevier, 2013.

[5] B.Ramesh and J.G.R.Sathiaseelan, "Support Vector Machine using Efficient Instant Selection for Micro Array Data Sets", IEEE, 2014.

[6] R. Sagayam, S.Srinivasan and 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.

[7] Stefano Ferilli, Floriana Esposito and Domenico Grieco, "Automatic Learning of Lingustic Resources for Stopword Removal and Stemming from Text", Procedia Computer Science 38 (2014) 116-123, Elsevier, 2014.

[8] Wahiba Ben Abdessalem Karaa,"A new stemmer to improve information retrieval", International Journal of Network Security & Its Applications, July 2013.

[9] Deepika Sharma, "Stemming Algorithms: A Comparative Study and their Analysis", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868, September 2012.

[10] Rupan Gupta and Anjali Ganesh Jivani, "Empirical Analysis of Affix Removal Stemmers", IJCTA, March- April 2014.

[11] Sandeep R.Sirsat, Vinay Chavan and Hemant S.Mahalle, "Strength and Accuracy Analysis of Affix Removal Stemming Algotithms", International Journal of Computer Science and Information Technologies, Vol. 4(2), 2013, 265-269.

[12] Giridhar N.S,Prema K.V and N.V Subba Reddy,"A Prospective Study of Stemming Algorithms for Web Text Mining", Ganpat University Journal of Engineering & Technology,Vol-1,Issue-1,Jan-Jun-2011.

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

[14] Prasenjit Majumder, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra and Kalyankumar Datta. "YASS: Yet another suffix stripper". ACM Transactions on Information Systems. Volume 25, Issue 4. 2007, Article No. 18.

[15] VenkatSudhakaraReddy.Ch and Hemavathi.D, "Information extraction using RDBMS and stemming algorithm", International Journal of Science and Research (IJSR), April 2014.

[16] C.Ramasubramanian and R.Ramya,"Effective Pre-processing activities in text mining using improved porter's stemming algorithm", International Journal of Advanced Research in Computer and Communication Engineering, December 2013.