International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 05 , May 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

DOCUMENT ANALYSIS WITH IMPROVED STEMMING ALGORITHM

Priti Shende[1], Prof. V. B. Kute[2]

[1,2] Department of Computer Science and Engineering,
Rashtrasant Tukadoji Maharaj Nagpur University, (India)

ABSTRACT

Online gathering of knowledge is getting more importance day by day. Data mining technique provides accurate results according to need. Choosing appropriate documents from large database is based on the frequency of words and their relation with one another. Frequency indicates the occurrence of a word in an article. Relation indicates the distance between two words. Different articles contain different forms of words. Therefore, words need to be converted into their stem form. Converting words into their stem form for calculating their frequencies leads to the incorrect words. In this paper, we propose a method which converts different forms of word in one group. A group is made after matching starting few letters of words. Those words are replaced by same word in that group having greater frequency. New length is calculated for again creating new group of words having same stem. New length equals to 60% of greatest length of word in a group. New groups are made after matching letters of words with one another equals to new length. Now all the words in new group are replaced by the word having greater frequency in new group.

Keywords: Comparison, Frequency, Lemmatization, Porter Stemming Algorithm, Stemming,

I. INTRODUCTION

Data mining technique is used for extraction of knowledge from large database. Data mining is applying methods like neural network, cluster analysis, genetic algorithms, decision rules and decision trees for extracting knowledge from large datasets. Large amount of raw data is collected in databases. But it is useless till we extract desired information from it. Various journals are available for different research areas. Research areas are vast and include variety of topics. Analyzing articles of journal for their areas of publications and quality of contents is very difficult task. Genealogy of research requires for thoroughly gathering of knowledge. Latest research includes new topics and phrases. A method is required to give such current topics from the journal. Comparison of articles on what basis is the big question. Computer doesn't understand of giving priority to articles. So the answer is comparison of articles can be done on the basis of frequency of words. Frequency indicates number of times a word present in an article. Articles contain different forms of words. Stemming method is used for converting words into their stem form. Now, all the words in each journal are in its stem form. Frequency of a word increased due to conversion of different forms of word to that single word. Article contains higher frequency of words has higher priority. Words are also related with one another on the basis of distance between them. Priority of articles can be determined according to word frequency, word relations and

user's query. Thus, it decreases load of database of handling different forms of words. Time also reduces for manipulating data. Representation of words and phrases graphically can give quick and better understanding.

II. RELATED WORK

Studying any field requires thorough knowledge. It refers to the research from the scratch until now. New research includes modifications in the topics and hence new phrases are added with the existing ones. New phrases of words also relate with other research fields. A new method for analyzing full text content of journal is determined in [1]. Co-citation analysis gives new topics of research and their relation with topics of other field if exists after calculating frequency of words and their inter-relation with one another. Visualization structure of words is available to get quick understanding. Following are the steps to determine word frequency.

2.1 Providing articles as input

Articles are taken as source material for analysis.
Articles are sorted according to volume, issue and paper.

2.2 Extracting main text and discarding unwanted characters or words

Analysis is done on main text of article. Take main text of article up to references. Each article has its own file for all the words it contains. This is a text for analyzing. Undesirable words from headline, keywords and authors need to be discarded.

2.3 Determining frequencies for each word and applying Porter Stemming algorithm

Calculate frequency of each word. Frequency determines occurrence of that word in an article. Each article has its separate file for frequency of words. Remove common words like we, do, from. Now apply basic Porter's stemming algorithm for converting each word in its stem form. Now, files contain stem form of all words.

2.4 Word pair formation

Determine frequencies of words. Word pairs are formed by measuring number of words between that two considered words. Minimum distance between those two words is considered in whole article. Number of times those two words appear by having some distance is their frequency.

2.5 Combining words of each article

All the words from each article are combined for their total calculation of frequency. Finally two files are generated. One file for word frequencies and second file for word pair frequencies.

2.6 Result in the form of graph

Force directed placement algorithm is used for visualization of network of words. Visualization structures are created according to word frequencies and word pair frequencies. So, it is easy to determine frequent topics and new phrases of that journal.

Porter's stemming algorithm converts word into its stem form after applying set of rules. Rules are applicable based on vowel-consonant pair. Rule is not applied when the vowel-consonant pair is zero. Porter's stemming algorithm often generates stem which is incomplete word. For example "probate" converted into "probat", this

is an incomplete word. Porter's stemming algorithm is modified by adding 31 rules as in [2]. It includes 15 additions, 11 replacements and 5 deletions. Stemmer correctness measures are overstemming, understemming, recall and precision. But they considered stemmer producing intelligible stems is the only performance measure. Meaningful stems can be produced after adding 31 rules as compare to original Porter's stemming algorithm. 30 words are taken and both the algorithms are implemented on them. Graph is plotted for performance comparison of generated stems. Addition of rules gives meaningful & complete stems in 93.3% cases of words which was 23.3% in original Porter's stemming algorithm. Thus error rate reduces from 76.7% to 6.7% according to graph in [2].

Another method of converting a word into its root form is Lemmatizing. Lemmatizing involves grouping of different forms of a word and considering one word as a base form that is lemma of a word. But it involves determining part of speech of a word in a sentence and understanding context. Stemming algorithm converts word into its stem form without considering context and parts of speech. So, stemming method is fast and easy to implement for retrieving documents. Less accuracy than lemmatization method doesn't matter for some application.

Errors in stemming depend on over stemming and under stemming. Over stemming indicates stemming of two words having different stems into same stem. Under stemming indicates stemming of two words having same stem in different stem. Algorithms have to take care of these two conditions.

Numbers of stemming algorithm are available. They are classified into truncating method, statistical method and mixed method. Truncating method includes algorithms like Lovins, Porters, Paice and Dawson. Statistical method includes N-gram stemmer, HMM and YASS. Mixed method includes inflectional and derivational method which contains Krovetz and Xerox, corpus based method and context sensitive method.

Truncating method is affix removal method. It means removing of prefixes and suffixes from word. Basic truncating algorithm keeps n letters of word and removes the rest. S- Stemmer makes plural form of word into singular form.

Paice stemmer takes care of deletion and substitution but having heavy algorithm. Dawson stemmer is fast in execution but having very complex algorithm. HMM stemmer is based on unsupervised method. Therefore, it is language independent method. For implementation it is a complex method. Krovetz stemmer act as pre-stemmer for other stemmers while implementation. It believes on meaning is necessary for removing suffix. It produces words not stems. It depends on dictionary and conservative. Xerox inflectional and derivational analyzer includes inflection and derivation. Inflectional database changes word into its base form which is different from its root word by tense, case, gender, number, person, voice and mood. Derivational database changes word to their stem form which is related with the root word by both form and semantics. Comparative study of all the algorithms of stemming is done on the basis of their advantages and limitations. No algorithm is totally perfect. But each of them satisfies one or other needs of stemming. Among these methods Porter's stemming algorithm is popular and produces best output as compared to other stemmers with less error rates suggested by [3].

Basic Porter stemming algorithm for removing suffixes from the word is given in [4]. It includes 5 steps containing various rules. It is a linear step algorithm. Rules are mostly depending on occurrence of vowel-consonant pair in a word. This is the important and main condition. It contains other conditions like stem ends with S, stem ends with double consonant, stem contains a vowel, stem ends with consonant-vowel-consonant

pair where the second C is not W, X or Y. These conditions are used in combine to determine whether the word is suitable for applying rule. If word is suitable for applying rules then it is converted into its stem form otherwise it is considered as a basic form of word and kept as it is.

Importance of Porter's stemming algorithm is suggested by [5]. It is widely referred and implemented for word conflation. Porter also generates rules of stemming for other languages. Snowball is a high level computer programming language produced for describing rules of Porter very clearly.

Porter's stemming algorithm is less in complexity for applying rules to remove suffix from the word as compare to Lovins' algorithm. According to Lovins' algorithm, there are more than 294 suffixes and these suffixes are removed according to 29 context sensitive rules. There is a risk of over-stemming. Thus, Porter stemming algorithm is getting more importance day by day.

Porter's stemming algorithm ignored many cases of suffixes for stemming [6]. More than 5000 suffixes such as -ativist, -ship, -ist, etc. are not handled. This definitely decreases the quality of documents retrieving. There are two categories of irregular forms. These are ignored by Porter. First form includes words like "bought" which is past participle of "buy". Dictionary is included containing list of such words. These words don't come under any rule. Directly dictionary is referred and their stem words are taken out. Second category includes words like –men which is plural form of –man. Rules are proposed for such words to convert them into their stem form. Thus an improved version of Porter's stemming is suggested in [6]. The new stemmer is evaluated using the Paice evaluation method. Error rate relative to truncation (ERRT) defined by Paice is used to calculate the overall accuracy of stemmers. Best stemmer should have less ERRT value as compare to other stemmers. New Porter, Porter, Paice and Lovins algorithms are implemented on word lists and it is found that ERRT value is least in case of New Porter. Retrieving several relevant documents by an information retrieving system is called recall. Retrieving less non-relevant documents is called precision. A good stemmer should have high values of recall and precision. New Porter stemmer has higher value of recall and precision than original Porter's stemmer algorithm. Thus, New Porter stemmer is better than original Porter's algorithm and other English stemmers.

Stemmers are limited to certain languages because of their rules. Statistical algorithms are used for information retrieval in absence of extensive linguistic resources for certain languages. Clustering based approach is used to convert different forms of word into their root form [7]. Information retrieval task divides into indexing and retrieval. In lexicon clustering, distance measuring technique between words is used to cluster different forms of a word into one cluster. K-means clustering is not suitable for this approach. Graph-theoretic clustering algorithms are used for detecting natural clusters in the data. Words are then stemmed into the word having central position in that cluster. These groups are represented using tree. Single-linkage, average-linkage and complete-linkage are three popular variants of graph-theoretic clustering. They are hierarchical in nature. But they differ in the criteria for similarity among groups.

In single-linkage method, two groups are similar if one of the members from each group having maximum similarity. In average-linkage cluster, mean similarity is defined between two groups of points by their mean similarity between points. In complete-linkage method, two groups are similar if one of the members from each group having minimum similarity.

In graph-theoretic clustering method, threshold value is inversely proportional to number of clusters. If threshold value increases then less numbers of big clusters are formed. If threshold value decreases then more

numbers of small clusters are formed. So, choosing appropriate threshold value is very significant in cluster based stemming method. Different queries can use to compare between algorithms. Information is available online in different languages but they are getting less attention. So, corpus based stemming proved good in information retrieving.

Information Retrieval model is suggested for finding desired information quickly from web by [8]. Data mining is very necessary for retrieving knowledge from web. Web mining includes large dataset but data mining refers to extraction of knowledge from large dataset.

Reference [9] suggested a language independent stemmer named as Single N-gram Stemming. Some N characters are same in each form of base word. Starting four characters of any word represents stem. Levenshtein distance is used for comparing results of Porter's stemmer and N-gram based stemmer. The Levenshtein distance is equal to number of deletions, insertion and substitution required for generating stem from word. Its result is not worse than Porter's linguistic Stemmer.

Summary of strength and similarity of four affix removal algorithms is given by [10]. These algorithms are Lovins, Paice, Porter and S-removal. Affix removal algorithms are most commonly used for stemming purpose. Correctness of stem is measured by closeness of converted stem to its root form. Another measure is to find how many semantically related words are assigned to same root word. Hamming distance function is used to find similarity among four affix removal stemming algorithms. Different algorithms are suggested so that it is easy to choose one of them according to need. Following are the six metrics to compare the strength of four stemmers.

a) The number of different forms of words assigned to one class is the measure of stronger stemmer.
b) Stronger stemmer gives larger index compression factor. In other words, total number of words should be larger than total number of stems. Percentage of index compression factor increases if total number of words increases and total number of stem decreases.
c) Stemmers don't remove characters from root word. But stronger stemmers also changes root words.
d) Stronger stemmers remove large number of characters from words in average to form stem. For example, in case of {effective, effect, effects} the root word is "effect". Mean= (4+0+1)/3 = 1.66. But this method does not measure transformations of stem endings.
e) So, new measure is mean modified hamming distance. Hamming distance is number of characters different at same position for equal length of words. The mean modified hamming distance of original words and stem is summation of number of characters different from original word as compare to stem for each word to number of original words. For example, in case of {try, tried, trying} the root word is "tri". Mean modified hamming distance is (1+2+4)/3 = 2.33.
f) Median modified Hamming distance is 2 in above case.

Suppose three stemmers stem three words. Similarity of stemmers is defined by number of words stem to the summation of difference in the characters at same position. Most often the stemmers don't change the word.

For example, in case of {exceed, demonstrated, arrival} one algorithm produces roots {excee, demonstrat, arriv} and another algorithm produces roots {excess, demonstrate, arrive}. The result is 3/ (2+1+1) = 0.75 as the measure for finding similarity between two algorithms. So, four algorithms can be compared pair wise. On the basis of mean and median modified Hamming distance, mean characters removed, compression factor, mean

conflation class size, word and stem different Paice stemmer is considered as the strongest stemmer in strength. Lovins is weaker than Paice but stronger than Porter. S-removal stemmer is the weakest in strength among these four stemmers [10]. One way to improve stemmer performance is to create one list of words for which stemmer fails. If the words exist in the document the word can be replaced by correct stem without applying any rule or statistics.

III. PROPOSED WORK

Stemming gives incomplete words. So, Instead of removing prefixes or suffixes words can be converted into one of the words among them having greater frequency. If initial few letters are matched then these words are considered in one group. Calculate new length equals to 60% of the greater length of word in a group. Words are again compared and their letters should match with one another equals to new length. Words having common letters up to new length make new group. Words are replaced by a word having greater frequency in new group. So, we get complete words. Results can also be compared for 50% and 70% of the greater length of word in a group.

IV. SUMMARY

Various algorithms have been studied. One algorithm is best in some field and the other algorithm is best in another field. Strength and similarities among algorithms gives their status. According to need different algorithms can be selected. Porter stemming algorithm is widely used. Document analysis using Porter's stemming algorithm with added 31 rules reduces errors but still gives incomplete words. Lemmatization produces better result but the method is complex. Therefore, Porter's stemming algorithm is selected for analyzing of articles. Documents analysis using improved Porter's stemming algorithm [2] and [6] can be done. Comparison of results can be done with Porter's stemming algorithm, improved Porter's stemming algorithm and proposed method.

REFERENCES

[1] Danilo Saft and Volker Nissen, "*Analysing full text content by means of flexible co-citation analysis inspired text mining method- exploring 15 years of JASSS articles*" *Int. J. Business Intelligence and Data Mining, Vol. 9, No. 1, 2014*

[2] K.K. Agbele, A.O. Adesina, N.A. Azeez , & A.P. Abidoye, "*Context-Aware Stemming algorithm for semantically related root words" in African Journal of Computing & ICT Vol 5. No. 4, June 2012*

[3] Ms. Anjali Ganesh Jivani, "*A comparative study of Stemming algorithms*" *in Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938*

[4] M.F.Porter, "*An algorithm for suffix stripping*" *Originally published in Program, Vo1. 4 no. 3, pp 130-137, July 1980.*

[5] Peter Willet, "*The Porter stemming algorithm: then and now*" *in electronic library and information systems, 40(3).pp. 219-223*

[6] Wahiba Ben Abdessalem Karaa ,"*A new stemmer to improve information retrieval*" *in International Journal of Network Security And Its Applications(IJNSA), Vol. 5, No. 4, July 2013*

[7] Prasenjit Majumder, Mandar Mitra, Swapnil K. Parui and Gobinda Kole , Pabitra Mitra and Kalyankumar Datta " *YASS: Yet Another Suffix Stripper*" *ACM transactions on information systems, vol. 25, no. 4, article 18, publication date: October 2007*

[8] Minyar Sassi Hidri and Amel Grissa Touzi, "*An Information Retrieval Model from World Wide Web based on Formal Concept Analysis*", *in international Arab Journal of e-technology, Vol. 2, No. 4, June 2012*

[9] B. P. Pande, Pawan Tamta, H. S. Dhami , "*Generation, Implementation and Appraisal of an N-gram based Stemming Algorithm*" *in press*

[10] William B. Frakes, Christopher J. Fox, "*Strength and similarity of affix removal stemming algorithm*" *in press*