Open access Journal International Journal of Emerging Trends in Science and Technology

A Study on Stemming Algorithms

Author
Anvitha Hegde[1], Mrs Savitha K Shetty[2]
[1]M.Tech, Dept. of ISE, MSRIT, Bangalore
Email: *anvitha.hegde@yahoo.in*
[2]Dept of ISE, MSRIT, Bangalore
Email: *savitha_ks@hotmail.com*

Abstract
Stemming is a process in in which different morphological variants of a word is transformed into its root form. A stemmer involves removing prefixes and suffixes of a word to reduce a word to its root form. This process is used in information retrieval systems and text applications in order to improve the efficiency of text retrieval.

Keywords: *over stemming, under stemming, inflections, derivational analysis, rule based stemmer*

INTRODUCTION

Stemming is a process in which different grammatical variations or deviations of a word are mapped to the root word. Stemming is used in query processing applications .while processing a query the stemming algorithm usually reduces the number of documents retrieved as it maps several terms to a single word. Thus accurate documents are displayed to the user. In the simplest form of stemming, all the root words are stored in a table and the table is queried to find a word which matched with the given word. Some algorithmsremove the suffixes oraffixes (prefix and suffix) before searching for the matching words. In certain stemming algorithms the suffixes are substituted then the matching words are searched. A stem dictionary is maintained and the root word id searched in the dictionary in some algorithms.

COMMON ERRORS IN STEMMING

Over stemming and under stemming are the most common problems of stemming. In over stemmingmorphological forms of a word which have different base meanings are mapped to the same root word. In under stemming word which have the same base meaning are mapped are mapped to the same root word.

COMMON STEMMING ALGORITHMS

Porters Stemmer

A framework known as snowball was designed by porter [1,2] based on which stemming algorithms suited for different languages can be developed. It is a rule based algorithm. Rules are applied on the words and if a rule is satisfied then the suffix is removed. It has over sixty rules. The error rate of Porters algorithm is very low.

Lovins Stemmer

This was one of the earliest stemming algorithms. This suffix stripping algorithm [3].It is a non-iterative algorithm and performs the lookup only once. A table whichstores the list of suffixes and rules is used in this algorithm. In a pass the loving stemmer removes the largest suffix found in the word. This algorithm is fast as it is notiterative; however it is limited to the number of suffixes found in the table.

Husk Stemmer

This is an iterative algorithm [4,5] and makes several passes. In each pass it apples a rule based

on the ending of the word. Deletions and replacements are made based on the rule applied. If no rule is found then the iteration stops. If there are three characters left in the word and the starting character is a vowel the iteration stops, if there are four characters left in a word and the starting letter is a consonant the iteration stops. However there is a possibility that words with different meanings are mapped to the same root word due to the replacements and deletion of characters.

Dawson Stemmer

This stemmer [6] is similar to the Lovins stemmer. It is a non-iterativealgorithm. It also maintain a table of suffixes similar to Lovins stemmer however the list of suffixes is much more comprehensivethan that of Lovins stemmer. The suffixes are stored in the table based on the lengths; the longest suffixes are stored first. Dawson stemmer is fast as it is non-iterative; however it is very complex on account of its comprehensive suffix list.

N gram method

This is a statistical method [7].It is not dependent on the underlying language gram is a set of n characters extracted from a continuous word. The entire text document is analysed .The variants of a root occur more frequently than the root; henceinverse document frequency can be applied on the document to identify them. This algorithm is very space consuming as it requires space to store the n grams and the indexes. The advantage of this algorithm is that it can be used in a variety of applications due to lack of language dependency.

HMM method

The hidden markov model (HMM) [8] is used in this method. No knowledge about the dataset is required in this method. A word is built as a result of transition between different states which occurs according to some probability function. The initial state is that of a stem and the final state is that of suffixes. The most likely path from the initial path to the final path is found in this method.

YASS Stemmer

Yet Another Suffix Striper [9] is a stemmer can be applied to different languages, it is language independent. Clustering such as hierarchical clustering is used in this method, distance measures are used and equivalence classes are created. The centroids of these classes are later determined.

Xerox Stemmer

Xerox stemmer can be used only for English documents as it is language dependent [10].A database which contains English words is used in this methodthis approach involves both inflectional and derivational analysis. After inflectional analysis the word are reduced to verbs, nouns and adjectives. The derivational analysis reduces the words to their stem words after removing the suffixes and affixes. The advantage of this method is that the resulting stem word is always a valid word which is present in the database. The disadvantage is that it is language dependent and depends on the words present in the database hence its efficiency depends on the comprehensiveness of the database. A lot of space is required to store the database. The efficiency of this stemmer increases if more words are stored in the database.

Minimum description length framework

This is an unsupervised learning method [11]. Grammar is developed from a set of heuristics. Minimum description analysis is done to ensure that the grammar produced from the set of heuristics is valid. Theadvantage of this method is that data is a compact form.

Hindi rule based stemmer for nouns

Even though Hindi is the national language of India there are not many stemming algorithms available for tithe Hindi rule based stemmer for nouns[12].This stemmer removes the suffixes of

words to get the stem word which is a noun. Rules were generated after researching the trending Hindi newspapers.

Rule based approach to extract inflectional and derivational words in Bengali

In this approach [13] the longest suffix is removed from the word . Two different hash tables were maintained which contained inflections. Both the tables were searched for the matching word,if the word was not found in the tables then it is considered as invalid.

Tamil stemmer

The recently developed stemmer for Tamil [14] used K means clustering. A rule based iterative algorithm [15] which removes the affixes to get the stem word which was developed using the snowball ball framework was tested on a tail dataset and yielded accurate results.

CONCLUSION

In this paper we discussed the different stemming algorithms, their advantages and disadvantages. There are basically two approaches in stemming. In the rule based approach rules are applied to find valid words and in the lexiconapproach, the database is searched to find a valid word. In both the approaches faces the problem of over stemming and under stemming.

ACKNOWLEDGMENT

I would like to thank Dr. Vijaya Kumar B P, Head of Department of Information Science Engineering, MSRIT and Mrs. Savitha K. Shetty, Assistant Professor,MSRIT for their valuable guidance.

REFERENCES

1. Porter M.F. “An algorithm for suffix stripping”.Program. 1980; 14, 130-137.
2. Porter M.F. “Snowball: A language for stemmingalgorithms”. 2001.
3. J. B. Lovins, *“Development of a stemmingalgorithm,”* Mechanical Translation and ComputerLinguistic., vol.11, no.1/2, pp. 22-31, 1968.
4. Paice Chris D. “Another stemmer”. ACM SIGIRForum, Vol. 24, No. 3. 1990, 56-61.
5. Paice Chris D. “An evaluation method for stemmingalgorithms”. Proceedings of the 17th annualinternational ACM SIGIR conference on Researchand development in information retrieval. 1994, 42-50.
6. Dawson, J. L. (1974); *Suffix Removal for Word Conflation*, Bulletin of the Association for Literary and Linguistic Computing, 2(3): 33–46
7. J. Mayfield and P. McNamee, “Single N-gram stemming”, *Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in InformationRetrieval*, 415-416, 2003.
8. M. Massimo and O. Nicola. “A Novel Method for Stemmer Generation based on Hidden Markov Models”,*Proceedings of the twelfth international conference onInformation and knowledge management*, 131-138, 2003.
9. 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.
10. Hull D. A. and Grefenstette,.“ A detailed analysis of English Stemming Algorithms”, XEROX Technical Report, http://www.xrce.xerox.
11. J. A. Goldsmith, “Unsupervised Learning of the Morphology of a Natural Language”, *ComputationalLinguistics*, MIT Press, 27(2):153-198, 2001.
12. Vishal Gupta, “Hindi Rule Based Stemmer for Nouns”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014.
13. Suprabhat Das, Pabitra Mitra, “A Rule-based Approach of Stemming for

Inflectional and Derivational Words in Bengali", Proceeding of the IEEE Students' Technology Symposium, PP.14-16, January, 2011.

14. M.Thangarasu. R.Manavalan, "Design and Development of Stemmer for Tamil Language: Cluster Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 7, July 2013.
15. Dhamodharan Rajalingam ,A Rule Based Iterative Affix Stripping Stemming Algorithm for Tamil", vol 132, PP-583-590, 2012.