*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 76-80*
*www.iosrjournals.org*

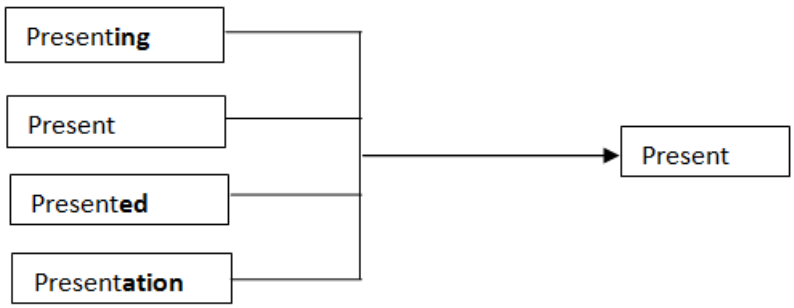
# A survey of Stemming Algorithms for Information Retrieval

Brajendra Singh Rajput[1], Dr. Nilay Khare[2]
[1,2] *(Computer Science &Engineering, Maulana Azad National Institute of Technology, India)*

***Abstract:****Now a day's text documents is advancing over internet, e-mails and web pages. As the use of internet is exponentially growing, the need of massive data storage is increasing. Normally many of the documents contain morphological variables, so stemming which is a preprocessing technique gives a mapping of different morphological variants of words into their base word called the stem. Stemming process is used in information retrieval as a way to improve retrieval performance based on the assumption that terms with the same stem usually have similar meaning. To do stemming operation on large data, we require normally more computation time and power, to cope up with the need to search for a particular word in the data. In this paper, various stemming algorithms are analyzed with the benefits and limitation of the recent stemming technique.*
***Keywords -*** *Information Retrieval, NLP, stemming technique, Decision based method, Statistical method.*

## I. Introduction

In Information Retrieval systems the main thing is to improve recall while keeping a good precision. A recall increasing method which can be useful for even the simplest Boolean retrieval systems is stemming. Information finder who is looking for texts say dogs is probably interested in the texts which consist of the term dog [6]. The capacity of the search database has increased in the last few years, so in order to meet the challenge of real time search NLP algorithms speed up required. Natural language texts typically consist of many different syntactic variants for example corrected, correct, correcting, correction, correctly, correctness, correctively, correctional, corrective, correctable (adjective), corrector (noun) all are derived word of root word correct [1].
The conventional approach used to extract data for some user query is to search the documents present in the corpus word by word for the given query. This approach is very time taking and it may leave some of the equivalent documents of equal nature. Thus to avoid these situations, Stemming has been extensively used in various Information Retrieval Systems to increase the extracting accuracy [4].
All documents which contain word with same stem as the query term are relevant, Stemming cut down the size of the feature set. In text mining, stemming can be viewed as clustering in pattern recognition, feature reducibility. In rule based reasoning, the main purpose is to choose maximum representative feature, dimensions base on similarity measurement [13].




The derived words present, presented, presentation and presenting are converted to root word present, through which not only retrieval performance improve but also storage can be optimized in some specific applications.

## II. Approaches Of Conflations

In order to perform stemming operation, we have to conflate a word to its different variants. Conflations approaches which are used in stemming algorithms are shown in figure 1. The conflation of words or Stemming can be executed in two ways, either manually using the kinds of regular expression or automatic. Automatic technique can be divided into four types namely affix removal, successor, table lookup and n gram. Affix removal can be further divided into two ways one is longest match and another is a simple removal [8].

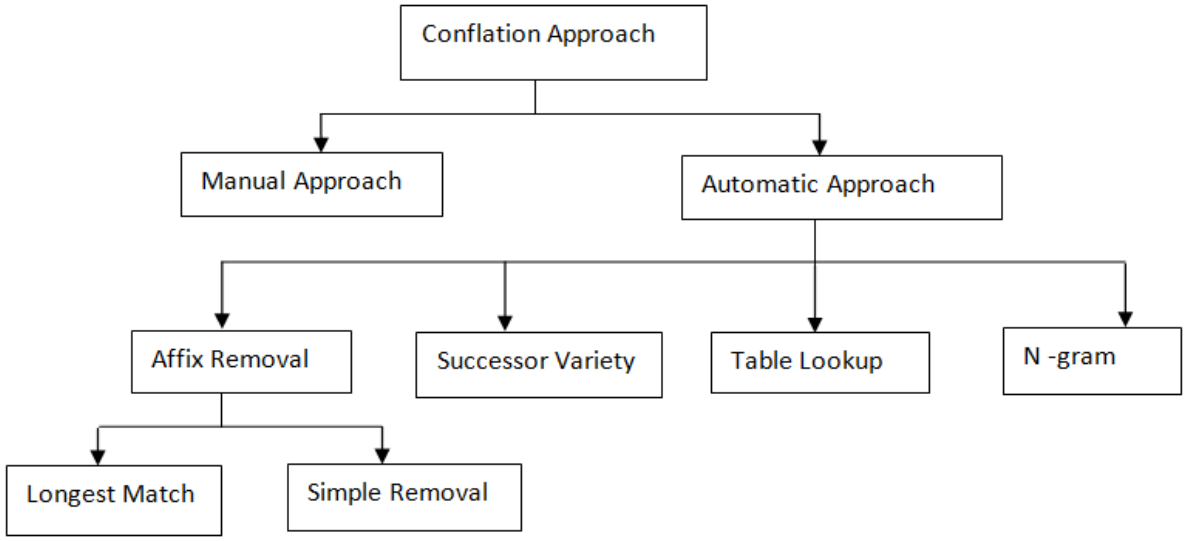



**Fig1:** Conflation Approach.

### 2.1 Affix Removal

The affix removal algorithms eliminate prefix or suffix from word in order to reduce word into common base. Most of stemmer used this type of approach for conflation. These algorithms depend on two principles one is iteration, which removes strings in each order class one at a time, starting at the end of a word and going towards its beginning. Not more than one match is allowed in a single order class. The suffix is added to a word in any random order, that is, there exist order classes of suffix. The longest match is second type in which within any given class of endings, if more than one ending gives a match then longest match should be eliminated [1].

### 2.2 Successor Variety

In successor variety method [12], frequencies of letter sequences in a body of text as the basis of stemming. The successor variety of a string is the number of different characters that follows it in word in some body of text. Consider text pattern which consists of the following terms for example, match, mean, mood, miasm, mobile .For estimating the successor variety (SV) for "machine" suppose, the following approach is used. The earliest letter of machine is 'm' which is accompany by **a, i, o, e** so successor variety of m is 4,for the next SV of machine we have to check that "ma" in machine is followed by which terms in the text body, so next SV of machine is 1 because t come next in match for machine. When this process is applied on a large body of text the successor variety of the substring of term will reduces as more character are added until a segment boundary is reached. So this idea is used to get the stem.

### 2.3 Table Lookup Method

Table lookup method is done by looking at the table where the term stems and their Corresponding stored. Term from queries and indexes could be stemmed by then a lookup table [6].If we use B-tree or hash table lookup then such would be fast, but there is a problem of storage overhead for such table.

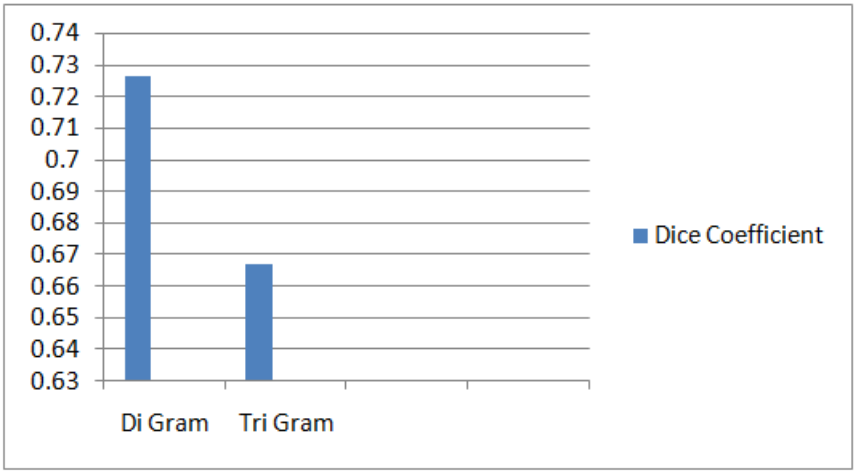
### 2.4 N-Gram Method

Another method of conflating the terms called shared diagram method given in 1974 by Adamson and Boreham [9]. The diagram is a pair of consecutive letters. Besides diagram, we can also use trigrams and Hence it is called n-gram method [10] .With this approach, pair of words are associated on the basis of unique diagram they hold both. For calculating this relationship, we use determines Dice's coefficient [8]. For example, the term Correction and Corrective can be broken into di-grams as follows.

| WORD | DI GRAMS | TRI GRAMS |
|---|---|---|
| Correction | *C,CO,OR,RR,RE,EC,CT,TI,IO,ON,N* | **C,*CO,COR,ORR,RRE,REC,ECT,CTI,TIO,ION,ON*,N** |
| Corrective | *C,CO,OR,RR,RE,EC,CT,TI,IV,VE,E* | **C,*CO,COR,ORR,RRE,REC,ECT,CTI,TIV,IVE,VE*,E** |
| A | 11 | 12 |
| B | 11 | 12 |
| C | 8 | 8 |
| Dice-Coeff. | 0.727 | 0.667 |

**Table 1** N – Grams (* denotes padding space)

Thus "Correction " has eleven digrams and twelve trigrams of which all are unique and " Corrective " also has eleven digram and twelve trigrams of which all are unique. The two words share eight unique digrams and trigrams.




Once the unique digrams and trigrams for the pair have been identified and counted, the similarity measure based on them can be calculated. The similarity measure is used Dice's coefficient, which is given as:
S = (2C)/ (A + B)

Where A is the number of unique N-gram in the First Word, B is the number of unique N-gram in the second word and C is the number of N-grams shared by A and B. For example, above Dice's coefficient would be equal (2 * 8) / (11 + 11) = 0.727 for Di gram and (2*8)/(12+12) = 0.667 for Tri grams. Such similarity measures are determined for all pairs of term in the database. Such similarity is computed for all the word pairs, they clustered as the groups. The value of the Dice coefficient gives you the hint that the stem for these pairs of words lies in the first 8 unique n-grams.

## III. Classification Of Stemmer

Basically Stemming algorithms can be classified into two types, Rule based and Statistical. Each type has its own ways to find for stem. Rule based stemmer encodes language specific rules, whereas statistical information from a large corpus of a given language to learn the morphology.

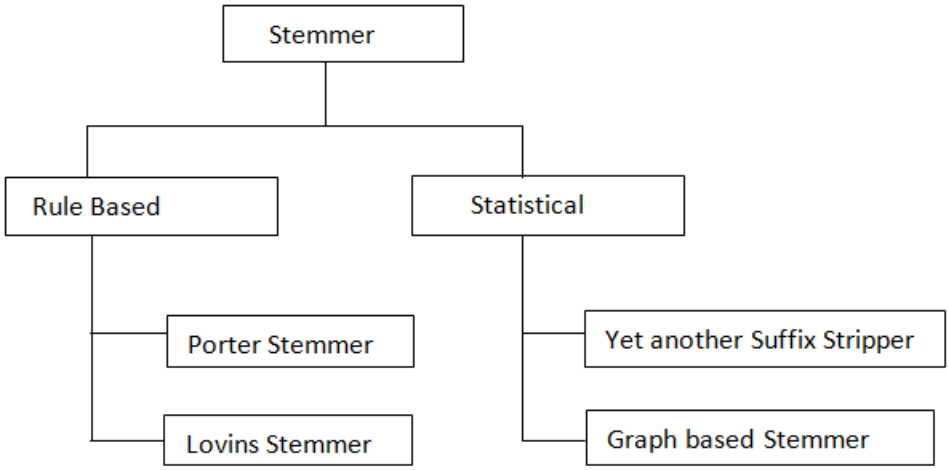



**Fig2:** Stemmer Classification.

### 3.1 Rule Based Stemmer

In a rule based approach language specific rules are planned and based on these regulations stemming is performed. In this approach various provision are specified for converting a word to its derivation stem, a list of all legitimate stem are given and there are some special rules which are used to handle the exceptional cases.

**3.1.1 Porter Stemmer:**

In standard Porter stemmer there are five steps and sixty conditions. There are many modifications of standard algorithms and its used for English document processing. General rule of removing suffix is given as:

**(Condition)S1 → S2**

Whenever condition is fulfilled suffix S1 is replaced by suffix S2. The order of consonants(C), vowel (V) and consonants (C) is counted as measure function (m) in porter stemmer. When the measuring function is greater than one, then only certain condition are applied [5].

**3.1.2 Lovins Stemmer**

In Lovins stemmer there are 29 conditions, 35 transformation rules and it perform a lookup on a table of 294 endings. Here stemming comprises of two phases [7].In the first phase, the stemming algorithm retrieves the stem of a word by removing its longest possible ending by matching these ending with the list of suffixes stored in computer and in the second phase spelling exception are handled. For example the word "absorption" is derived from the stem "absorpt" and "absorbing" is derived from the stem "absorb". The problem with the spelling exception arises in the above case when we try to match the two words "absorpt" and "absorb". Such exceptions are handled very carefully by introducing recording and partial matching techniques in the stemmer as post stemming procedures.

Rule dependent stemmer is fast in nature means calculation time used to find a stem is less. The retrieval result for English by using a rule dependent stemmer is reasonable, but the problem associated with rule based is one need to have extensive language expertise to make them.

**3.2 Statistical Stemmer**

Statistical Stemmer is good alternative to rule based stemmer and does not involve language expertise. They use statistical information from a large corpus of a given language to learn morphology. Statistical language processing has been successfully used to improve the performance of information retrieval systems in the absence of extensive linguistic resources for some language.

**3.2.1 Yet Another Suffix Stripper (YASS)**

Yet another suffix stripper is one of statistical based language independent stemmer and its performance can be compared with both rule base stemmer in term of average precision. In this method a set of string distance measure is used. The string distance measure is used to check the similarity between the two words by calculating string the distance between two strings. The distance function maps a pair of string a and b to a real number r, where a smaller value of r indicates greater similarity between a and b. The main reason for estimating this distance is to find the longest matching prefix [4].

**3.2.2 Graph Based Stemmer (GRAS)**

GRAS is a graph based language independent stemmer for information retrieval. Extracting effectiveness, simplification and low computation cost are the features of GRAS. In GRAS [10], first we look for long common prefix amongst the word pair available in the document set. Suppose two word pair $W_1=P*S_1$ and $W_2=P*S_2$ where P is the longest common prefix between W1 and W2.The suffix pair $S_1$ & $S_2$ should be valid suffix if other word pairs also have a common initial part followed by these suffixes such that $W'_1 = P' *S_1$ & $W'_2 = P'* S_2$ Then, $S_1$ & $S_2$ is the pair of candidate suffix if large number of word pairs is of this form.

Then look for pairs that are morphological related if they share a non-empty common prefix. The suffix pair is a legal candidate suffix pair. Using a Graph we model word relationships where nodes represent the words and edges are used to attach the related words. Normally in GRAS Pivot is a node which is associated by edges to an other nodes. In the last step, a word which is connected to a pivot is put in the same class as the pivot if it shares common neighbors with the pivot.

## IV. Stemming Error

There are fundamentally two kinds of fault in stemming algorithms one is over stemming and another is under stemming [3]. Over stemming occurs when two words which have dissimilar root word are changed to the identical base term, which is also identified as a false positive. In under stemming two words which have similar root are not stemmed to the same base term, which is also called as false negative. Paice [11] has demonstrated that light stemmer decreases the over stemming but increases the under stemming errors. On the other side heavy stemmer reduces the under stemming error while increasing the over stemming errors.

## V. Conclusion

We studied a variety of stemming methods and got to know that stemming appreciably increases the retrieval results for both rule dependent and statistical approach. It is also useful in reducing the size of index files and feature set or attribute as the number of words to be indexed are reduced to common forms called stems. The performance of statistical stemmers is far superior to some well-known rule-based stemmers but time consuming. Rule dependent stemmer like porter stemmer is good choice for English document processing but its language dependent.

## References

[1]. Sandeep R. Sirsat, Dr. Vinay Chavan and Dr. Hemant S. Mahalle, Strength and Accuracy Analysis of Affix Removal Stemming Algorithms, International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 265 - 269.
[2]. S.P.Ruba Rani, B.Ramesh, M.Anusha and Dr. J.G.R.Sathiaseelan, Evaluation of Stemming Techniques for Text Classification ,International Journal of Computer Science and Mobile Computing, Vol.4 Issue.3, March- 2015, pg. 165-171
[3]. Ms. Anjali Ganesh Jivani, A Comparative Study of Stemming Algorithms, International Journal Comp. Tech. Appl.(IJCTA) 2011, Vol 2 (6), 1930-1938 ISSN:2229-6093
[4]. Deepika Sharma, Stemming Algorithms: A Comparative Study and their Analysis, International Journal of Applied Information Systems (IJAIS) Foundation of Computer Science, FCS, New York, USA September 2012 ISSN : 2249-0868 Volume 4– No.3
[5]. Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3):130–137
[6]. Wessel Kraaij and Renee ´ Pohlmann,Porter's stemming algorithm for Dutch,UPLIFT (Utrecht Project: Linguistic Information for Free Text retrieval) is sponsored by the NBBI,Philips Research, the Foundation for Language Technology.
[7]. Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linquistics, 11:22–31
[8]. WB Frakes, 1992,"Stemming Algorithm ", in "Information Retrieval Data Structures and Algorithm",Chapter 8, page 132-139.
[9]. G. Adamson and J. Boreham 1974. "The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles," Information Storage and Retrieval, 10,253-60.
[10]. JH Paik, Mandar Mitra, Swapan K. Parui, Kalervo Jarvelin, "GRAS ,An effective and efficient stemming algorithm for information retrieval", ACM Transaction on Information System Volume 29 Issue 4, December 2011, Chapter 19, page 20-24
[11]. Paice Chris D.,Another stemmer, ACM SIGIR Forum, Volume 24, No. 3. 1990, 56-61.
[12]. M. Hafer and S. Weiss 1974. "Word Segmentation by Letter Successor Varieties," Information Storage and Retrieval, 10, 371-85
[13]. Narayan L. Bhamidipati and Sankar K. Pal, Stemming via distribution based word segregation for classification and retrival,IEEE Transaction on system,man,and cybernetics – partB cybernetics. Vol 37 ,no2 april 2007.