ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

PREPROCESSING IN INFORMATION RETRIEVAL

R.Rengaraj alias Muralidharan#,C. CaslinPenisha*, T.V. Charulatha*, P.Divyarani*, K.Yamuna*

Information Technology,SaranathanCollege of Engineering, Trichy.
*Final Year , #Assistant Professor,

1. ABSTRACT

Text categorization is a fundamental task in the document processing, it is automated to handle enormous streams of documents in electronic form. Major difficulty in handling documents is thepresence of morphological variant, homographs ,and textual errors such as spelling ,grammatical errors .In order to handle such situation preprocessing technique is used. Stemming is one of the major technique used in pre-processing , it conflate the variant forms of a word into a common form .The technique discussed in this paper are Stop words removal ,Case folding and Affix Stripping. These technique further reduces the size of the text in the input file

2. INTRODUCTION

The main aim of information Retrieval (IR) is to satisfy the user requirement and produce an readable automated summary.Automatic text summarization plays an important role in information industry, especially dueto the exponential growth of data in recent years. Stemming technique conflate the variant words into common representation. Has an excellent trade –off between speed readability and accuracy .The availability of search-engines enable us to retrieve the colossal amount of information. Many words in the documents often have morphological variants .So before performingsummarization,the stemming techniques are applied on the target data set to reduce the size of the data set which will increase the effectiveness of IR System. In this paper, surveys of stemming techniques also presented.

3.RELATED WORKS

Jivanistatesthe that rule based approach may not always give correct output and the stems generated may not always be correct words.The problem of over stemming and under stemming can be reduced only if the syntax as well as the semantics of the Words and their POS(point of scale) is taken intoconsideration .Since However no perfect stemmer has been designed so far to match all the requirements[1]. Sharma statesthatthe performance of statistical stemmers is far superior to some well-known rule-based stemmers and among statistical based stemmers GRAS (GRAPH BASED algorithm[2].Giridhar N S discussed about varies algorithms most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. In linguistic morphology ,stemming is the process for reducing inflected words to their stemthetechnique used to solve the problem of spellingexceptions and performs a number of transformations based on the letterswithin the stem..[3]. LailaKhreisatpoints out the results of classifying text documents using the N-gram frequency statistics technique employing a dissimilarity measure.N-gram text classification using the Dice measure outperforms classification.It calculates rank-order statistic for two profiles by measuring thedifference in the positions of an N-gram in two different profiles. For each N-gram in the document profile, search for the N-gram in the class profile and calculate thedifference between their positions.The N-gram method is language independent and works well in the case of noisy-text (text that contains typographical errors). Tri-grams for text classificationis used. The trigramsof a string or token is a set of continuous 3-letter slices of the string [4].Vishalrefers that generally to the process of extractinginteresting and non-trivial information and knowledgefrom unstructured text. Text mining is a younginterdisciplinary field which draws on informationretrieval, data mining, machine learning, statistics and computational linguistics. As most information (over80%) is stored as text, text mining is believed to have ahigh commercial potential value. Knowledge may bediscovered from many sources of information, yet,unstructured texts remain the largest readily availablesource of knowledge[6]. Parmarpresented a paper in which, depends on the key-phrases extracted by the system and many other features extracted from the document to get the text summary most related words are added to the summary she proposed algorithm that strengthen and reduce time taken for execution of process. Take multiple document into consideration dataset document for each document from dataset D, with set of team T and Sentence S.. All have opted text mining approach. there can be graph mining, multiview leading . The result will

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

be Top-N keywords extracted from multiple dataset.

4.PROPOSED SYSTEM

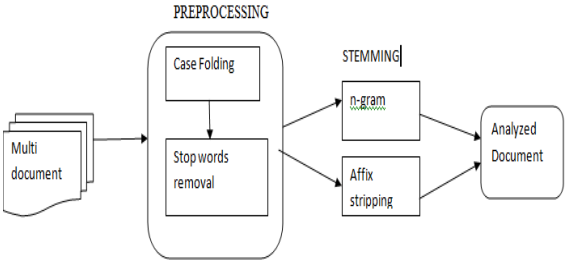

Fig1:Process involved in IR

4.1 Multi-document

Information from different news articleare gatheredfor extraction the relevant information from the document

4.2 PREPROCESSING

These are the two steps involved in the preprocessing it is on statistical natural language processing tool which filter out stop words and generate stem words, by avoiding the inflectional term .Each of the words in the documents were represented as vector-space model, the number of words would have been too high for the text summarization algorithm. Preprocessing methods that greatly reduce the number of words in the document .This convert the unstructured text into structured format and removes the stop words, and store the result in a table and extract the important key phrases in the text by implementation.In preprocessing stage case folding and stop words removal are done to reduce the size and increase the efficiency. The different stages in preprocessing are elaborated below

4.2.1 Case Folding

It is used toconvert all the characters into same case format, either the **upper-case** or **the lower-case** format.example:“act”, “Act”, “aCt”, “ACt”, “acT” converted into standard lower-case format “act”.

4.2.2 Stop Words Removal

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words in the document.Stop wordsare**"single set of words".** It means different things to different applications. For example, in some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. It createsan filter for words that need to be the candidate keywords.

The steps involved in stop words removal are

Step 1: The text document is tokenized and individual words are stored in array.

Step 2: A single stop word is read from stop word list.

Step 3: The stop word is compared to target text in form of array using sequential search technique.

Step 4: If it matches, the word in array is removed , and the comparison is continued till length of array.

Step 5: After removal of stop word completely, another stop word is read from stop word list and again algorithm follows step 2. The algorithm runs continuously until all the stop words are compared.

Step 6: Resultant text devoid of stop words is displayed.

***Pseudocode*:**It gives an algorithm for stop words removal.

Data: Document containing text

Result: Document with stop words removed

begin

Initialize:

words checked = “ ”

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

Perform:

```
for each d in document do
for each word in d do
result =callDFA STOPWORD(word)
if result == false then
//append because its not a stop word
words checked.append(word)
        end
end
return words checked
        end
end
```

Advantage of the algorithm isit helps in developing meaningful domain of our choice.

Disadvantageisin some search engine several common words get ignored.

4.3STEMMING

It is the process of conflating the variant forms of a word into a common representation .The fruitful usage of stemmer, is that morphological variants of words are semantically related. It improves the performance of the system.Most stemmers today, do not always output root words. Taking note of the fact that linguistic correctness of the stems may become critical to effective retrieval in future, design of a root-word stemmer has been proposed and used .

Affix Stripping Algorithmand N-gram Algorithm are used for stemming

4.3.1Affix Stripping Algorithm

The affix refers to either a prefix and suffix. In addition to dealing with suffixes, several approaches also attempt to remove common prefixes. Inflectional affix algorithm shouldnot affect the basic meaning of the stem and it is stemmed without risk of losing too much information.It reduces the number of terms in IR and reduce the size and complexity of data.
For example word *indefinitely* the leading "in" is a prefix that can be removed.

Algorithm Rules

To present the suffix stripping algorithm in its entirety we will need a fewdefinitions.

The `condition' part may also contain the following:

- *S - the stem ends with S (and similarly for the other letters).
- *v* - the stem contains a vowel.
- *d - the stem ends with a double consonant (e.g. -TT, -SS).
- *o - the stem endswhere the second c is not W, X or Y (e.g. -WIL, -HOP).

The algorithm is careful not to remove a suffix when the stem is too short,the length of the stem being given by its measure, m. There is no linguisticbasis for this approach. It was merely observed that m could be used quiteeffectively to help decide whether or not it was wise to take off a suffix.

Advantage of the systems are:

It is simpler to maintainbecause it donot rely on a lookup table and it depends on root form relationship.
Maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Overlapping rule does not occur in affix stripping.

Limitations are :

Linguistically incorrect stems:Some stems which are generated by this algorithm are not linguisticallycorrect. This maynot be a problem for unique and consistent group of words, butfor the identical stem this is not semantically correct and this will result in retrieval error
Homographs: Words which are spelled identically but nevertheless have a differentmeaning .It does not have access to, word categories,and the different types of wordsare not distinguished
Irregular verb: The verbs exhibits irregularities in the formation of past tense, past participleare almost completely irregular. Example:Break,Broke,Broken are completely irregular.

4.3. 2 N-gram

This Algorithm is extensively used in text mining and naturallanguage processing tasks. It is an N-character slice of a string. The N-gram method is language independent and works well in the case of noisy-textThey are basically a set of co-occuring words within a the given window and when computing the n-grams you typically move one

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

word forward .For example, "The dog barks at the cat". If N=2 (bigrams) then N-Gram would be

- the dog
- dog barks
- barks at
- at the
- the cat

The value of N=2 ,we have 5n-Grams in this case. we moved from the->dog to dog->barks to barks->at, etc, essentially moving one word forward to generate the next bigram .If the value N=3(trigrams)

- the dog barks
- dog barks at
- barks at the
- at the moon

In this we have 4 n-grams. When **N=1**, this is referred to as **unigrams** and this is essentially the individual words in a sentence.When **N=2 bigrams** and when **N=3** this is called **trigrams**.WhenN>3 this is usually referred to as four grams or five grams and so on.

Formula:

$$Ngrams_K = X - (N - 1)$$

K- Given Sentence
X=Num of words in a given sentence K
N=value of N (N=1,2,3…..)

Steps for Generating N-Gram

1) Split the text into tokens consisting only of letters. All digits are removed.
2) Compute all possible N-grams.
3) Compute the frequency of occurrence of each N-gram.
4) Sort the N-grams according to their frequencies from most frequent to leastfrequent. Discard the frequencies.
5) This gives us the N-gram profile for a document. For training class documents, N-gram profiles were saved in text files.

N-Gram Algorithm

int N, String sent, List ngramList

```
{
String[]tokens=sent.split("\\s+"); //split
sentence into tokens
//GENERATE THE N-GRAMS
for(int k=0;k<(tokens.length N+1);k++)
{
String s="";
int start=k;
int end=k+N;
for(int j=start; j<end;j++)
{
s=s+""+tokens[j];
}
//Add n-gram to a list
ngramList.add(s);
}
}//End of method
```

Advantage are:N-gram Algorithm is used for tasks such as spelling correction, word breaking and text summarization. For developing features for supervised MachineLearningmodel.N-gramprofiles provides a simple and reliable way to categorize documents.

***Limitation are*:**The*n*-gram models are not designed to l linguistic knowledge**.** In n-gram there is a lack of explicit representation of long range dependency.

The Sample output taken in various stages of the proposed system are presented below from fig 2 to fig6

Fig 2.Home Page

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 3, Special Issue 2, April 2017

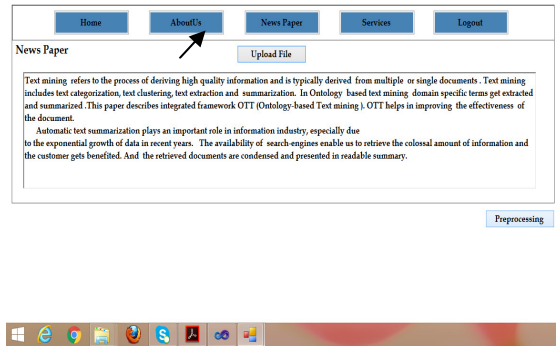

Fig3 .Uploading file content

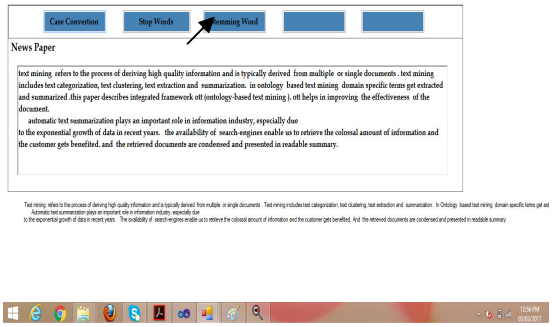

Fig 4: CaseConversion

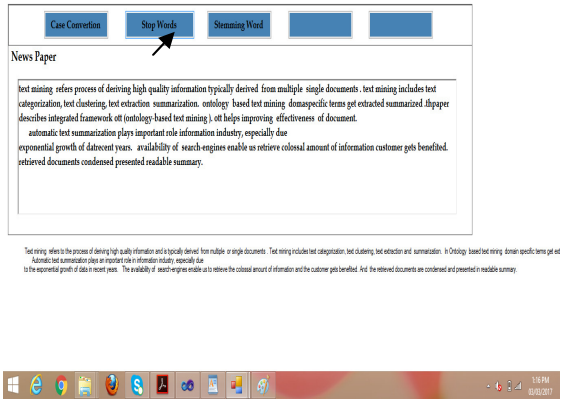

Fig 5:Stop words removal

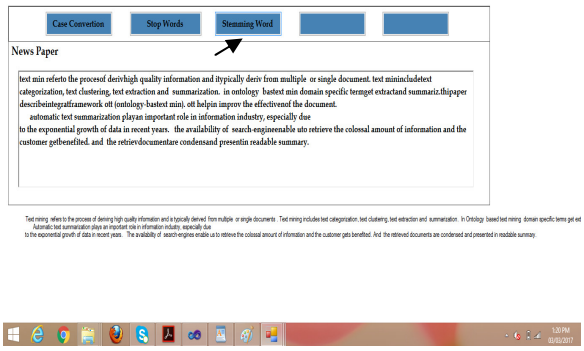

Fig 6: Stemming

5. CONCLUSION

This paper discusses the various preprocessing techniques used in text summarization for wide varietyof domains. Stemming can be effectively used in natural language processing for removing noisy data.It increases the retrieval result for both rule based and statistical approach and the index file size gets reduced . Lot of similarities lies between different Stemming algorithm.Each algorithmhas its own pros and cons none of them gives cent percent output since it can be applied to Text Mining applications.

6. REFERENCES

[1] Ms. AnjaliGaneshJivani:A Comparative Study of Stemming Algorithms"IJCTAJ.Comp.tec.ApplVol 2 (6), 1930-1938 IJCTA | NOV-DEC 2011.www.ijcta.com

[2] Deepika Sharma [ME CSE] "Stemming Algorithms: A Comparative Study and their Analysis"Foundation of Computer Science FCS, New York, USAVolume 4– No.3, September 2012 – www.ijais.org

[3] Giridhar N S, Prema K.V, N .V Subba Reddy "A Prospective Study of Stemming Algorithms for Web Text Mining.Department of CSE,M.I.TManipal University, Manipal, Karnataka, India 1. VOL.-1, ISSUE-1, JAN-JUN-2011

[4]LailaKhreisat "Arabic Text Classification Using N-Gram Frequency StatisticsA Comparative Study" A Comparative Study,Proceedings of the

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

2006 International Conference on Data Mining (DMIN'06): June 26 - 29 2006, Las Vegas, USA.

[5] ParmarParesh B. "A Survey Paper on Mining Keywords Using Text Summarization Extraction System for Summary Generation over Multiple Documents" International Journal of Science and Research (IJSR) ISSN (Online) Index Copernicus Value (2015): 78.96

[6] Vishal Gupta"A Survey of Text Mining Techniques andApplications" Journal of emerging technologies in web intelligence , vol. 1, no. 1, AUG 2009.

