ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

Implementation on Morphological Text Mining

Chitra Kapoor[1], Ms Bhawna Mallick[2]

Department of IT[1], Head, Department of CSE[2]

Galgotias College of Engineering and Technology, Greater Noida, India

ABSTRACT: Text mining is extracting useful or relevant information from a textual file. The extraction of information works on search engine or through a tool. Text mining is used for text retrieval and categorization. This paper focus on every word of a textual document whose frequency or occurrence in a textual document is high and those words has its own meaning i.e. they are stand-alone words. So we have developed a tool which will extract the details of textual data for a user in summarized form and the information subsequently is sufficient to know about a textual file.

Keywords: Text Mining, Morphology, n-gram, categorization, information extraction, text retrieval.

I. INTRODUCTION

Text mining — also called intelligent text analysis, text data mining, or knowledge discovery in text — uncovers previously invisible patterns in existing resources[1].Text mining is branch of data mining. Data mining is used to extract useful information or pattern from a large database and later can be transformed into usable form. Data mining has various branches some are text mining, pattern mining, web mining etc. However, our study is limited to text mining. Text mining is extracting some relevant, useful and important information from a textual document and the objective of text mining is to know the pattern and association between various texts. There are various tasks in text mining. These are text categorization, clustering, information extraction, text retrieval etc. It not only helps to find important text but also find frequency of the text. Text mining also tells the relationship between the text. There is difference between information extraction and text retrieval i.e. information extraction is to extract the piece of information from unstructured document whereas text retrieval is done before the information extraction i.e. it is the branch of information extraction and in this the important information is in the form of text. All the text which is achieved from the result will be considered and using this process the person will extract the desired information. Text retrieval is collecting all text from textual document and then analyzes the result for the details.

Text mining is used in various applications like security, biomedical, commercial and web. The proposed work focus on text mining which will help the user to get the summarized information of the textual document.

II. RELATED WORK

Sangno lee[et.al] has proposed empirical comparison of four text mining methods. It was limited to the review of statistical approach, and does not extend to the syntactical and morphological approach of natural language[6]. Peter Náther proposed N-gram based text categorization which has found a method for the text clusterization. It is a method, which helps us to build categories from the set of documents, without previous knowledge of given texts[4].R via the tm package offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R[5]. Suneetha Manne[et.al] has proposed a novel approach for text categorization of unorganized data based with information extraction proposed that information extraction is the main function in text mining which applies natural language processing (NLP) techniques to extract pieces of information (such as name, date, or affiliation), thus giving the document some structure. But it becomes a complex task when dealing with files of multiple formats like *html* files, *pdfs*, *docs etc.*, as they involve various media and graphical items[3].

III. PROPOSED TECHNIQUES

We have studied various text mining techniques like n-gram technique, only meaningful n-grams are retrieved from the textual document, text categorization, information extraction, Text retrieval without using the package and coding. Text

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

retrieval tool has been designed using the above techniques which will be helpful for future use to understand the document much better without reading the document. The related terminology is as follows :-

A. MORPHOLOGY

Morph means 'form'. Morphology is study of word form or structure and this structure is in form of morpheme. Morpheme is smallest unit which has its own meaning i.e. stand-alone word.

B. N-gram TECHNIQUE

N-gram is a sequence of text (where n is equal or greater than 2). N-gram can be meaningful or non-meaningful text.Meaningful text subject to those which have some meaning i.e. remove stop words (a, an, with ….. etc). RGui[2] shows n-gram technique very effectively.

C. TEXT MINING FUNCTIONS

Information Extraction and Text categorization are the tasks which are done by text mining.

1) *Information Extraction:* Information Extraction is an important function in text mining. It is extracting summarized information in form of short text from a textual data.

2) *Text Categorization:* Text Categorization can be defined as a content-based assignment of one or more predefined categories to free texts[4] .i.e. main-category has its sub-category and every text has its own category.

D. TEXT RETRIEVAL APPROACH

Our work focus on a tool which will support information extraction. This tool is known as "TEXT RETRIEVAL TOOL".

Its function is to extract the relevant information from a textual document. This will remove all the stop words and extract meaningful words. According to user's own need he can remove any word which he don't want to include in the result. If a textual document is not known to a user then also a user will be able to know the detail of document through short texts. With every text its occurrence in a document will also be extracted which means how many times a word occurred in a document and the highest occurred texts which are known as 'keywords' i.e highest frequency words are calculated separately which focus on the topic of the document.

E. CATEGIRIZATION of TEXT THROUGH TEXT RETRIEVAL TOOL

Categorization of text can be based on several ways. But in our tool categorization of text is on the basis of text. The text which have same number of frequency will be shown together therefore, we can say more the frequency of the text the more information can be retrieved. There are several ways to retrieve the information in this tool like all the important text of textual document, all the keywords i.e. most frequent words of textual document and there are some other option like more than three frequency words and more than four frequency words. In figure 1 the working of text retrieval tool is shown.

First the textual document is inserted in text retrieval tool through this tool some important keywords and other relevant information is produced as result. These information is enough to describe the document.

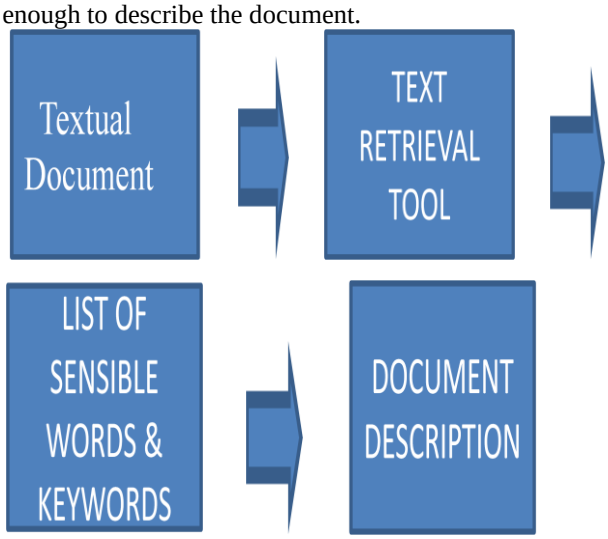

Fig. 1 Working of Text Retrieval tool

IV. HOW TEXT RETRIEVAL TOOL DIFFERENT FROM RGUI

The same work can be done in RGui. Firstly, in this a user has to install the package first and every time he has to load the package and then start the work, in this environment a user has to do the coding for relevant result this will be time consuming. The result of RGui is shown in figure 2 but in Text Retrieval tool no need to load the tool again and again. Secondly, in Text Retrieval tool along with all meaningful word's frequency it also calculate, another advantage of text retrieval tool is that no need of coding so no wastage of time and anyone can use this tool. The user have to insert the textual document and submit it so that the document will be saved and then by browse button the results can be shown. It also contain automatic clock. Text Retrieval tool shown in figure 3. There are many options in this tool like REFRESH button this will refresh the whole tool to start the new work; SUBMIT button help the user to submit the document for future use; RESULT button is used to search the saved document with the help of document name, the whole detail

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

about the document can be shown; DELETE button is used to delete the document which have been saved by the user; EXIT button helps the user to exit from this tool.

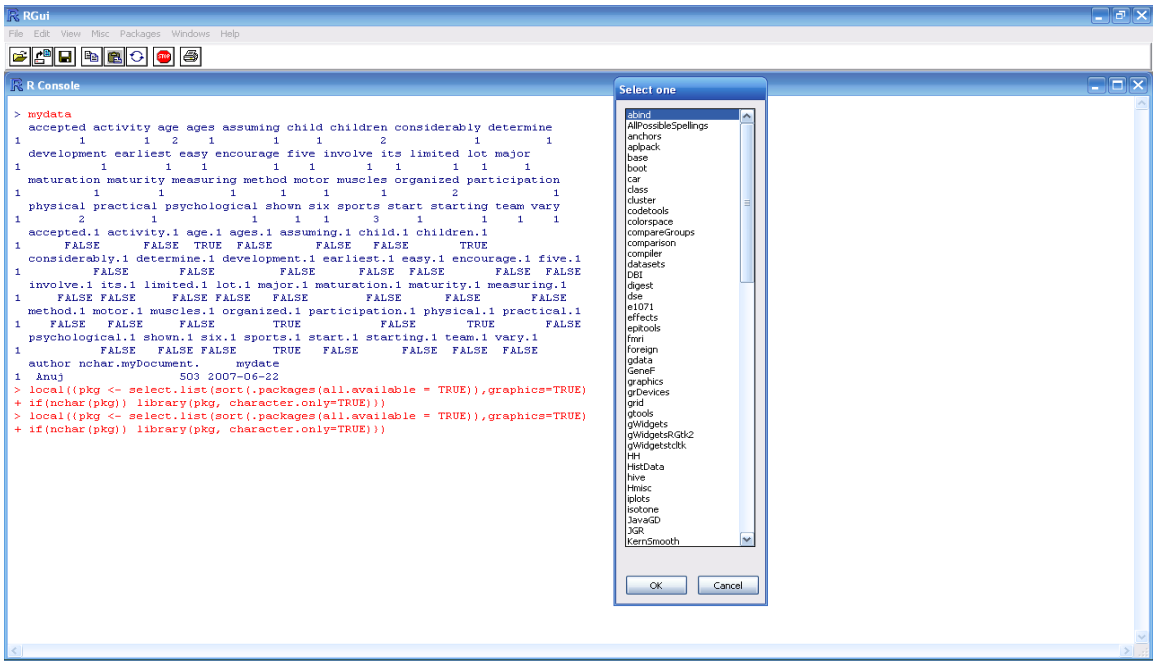

Fig. 2 Result obtained by using RGui Tool

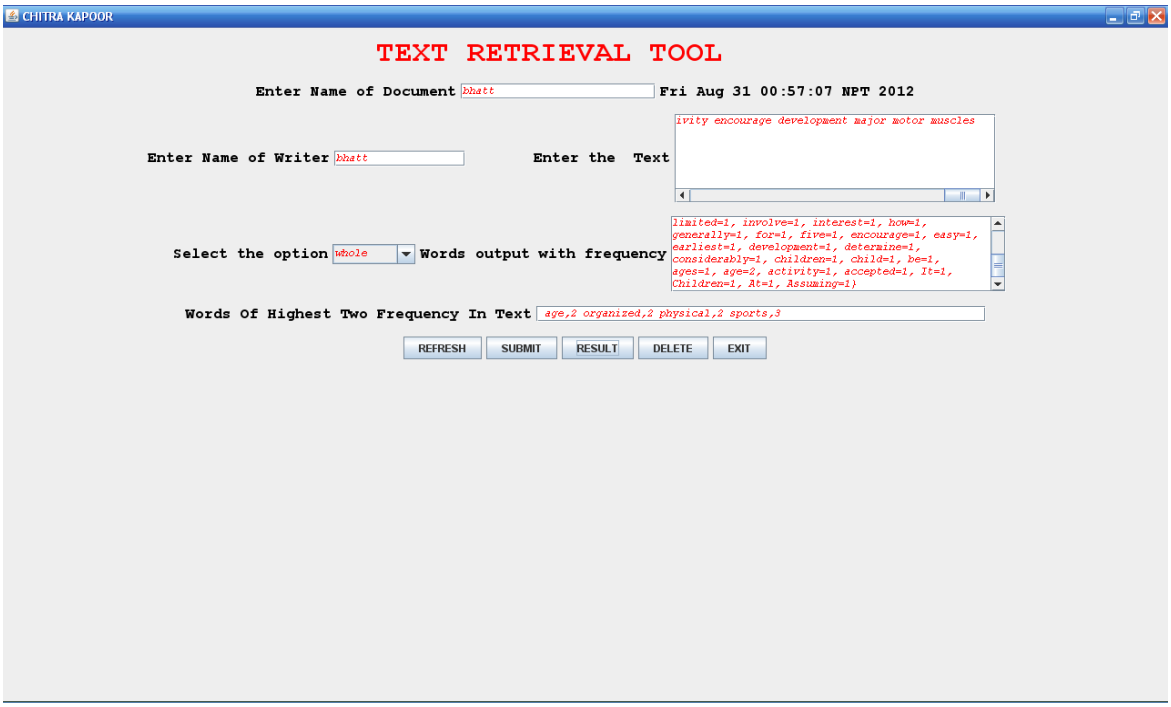

Fig. 3 Result obtained by proposed Text Retrieval tool

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

V. APPLICATIONS

This tool can be used in various purposes like

A. Survey
B. Measuring customer preferences
C. Related articles
D. Monitoring public opinion like in social networking sites etc.
E. In software application to improve the results

VI. CONCLUSION

In this paper we have proposed a tool which is Text Retrieval tool. Its function is to calculate all important meaningful words. It calculates frequency of all meaningful words As well as keywords. The user can get the idea about the document by getting the most often used words in the text. One has to only enter the document name, to get the relevant detail regarding the document. In this paper a new tool is introduced which will help the user to know about the textual document. It will give the keywords with its frequency (occurrence of word) in the document . It only works on one parameter which is text retrieval. In the future, this work can be extended for various other parameters like clustering, pattern recognition, association analysis etc. so that user can get more accurate information using this tool.

REFERENCES

[1]. Atika Mustafa, Ali Akbar, and Ahmer Sultan,"Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", *International Journal of Multimedia and Ubiquitous Engineering*,2011

[2]. G. Koteswara Rao and Shubhamoy Dey , "DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH" ,*International Journal of Managing Information Technology (IJMIT)* Vol.3, No.3, August 2011

[3]. Suneetha Manne and Dr. S. sameen Fatima, "A Novel Approach for Text Categorization of Unorganized data based with Information Extraction", *International Journal on Computer Science and Engineering*,2011.

[4]. Peter Nather, "N-gram based Text categorization"

[5]. Ingo Feinerer, Kurt Hornik and David Meyer, "Text Mining Infrastructure in R", *Journal of Statistical Software*,March 2008, Volume 25, Issue 5.

[6]. Sangno lee, Jaeki Song, Yongjin Kim, "An Empirical Comparison of Four Text Mining Methods*" Journal *of Computer Information Systems*, 2010

Biography

Chitra Kapoor has completed B.E in Information Technlogy from B.S Anangpuria Institute of Technology and Management, Faridabad, India, in 2010 and pursuing M.Tech from Galgotias College of Engineering and Technology, Greater Noida, India.