www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 4 April 2016, Page No. 16253-16256

Multi-Document Summarization Using TF-IDF Algorithm.

Prof. Amit Savyanavar, Bhakti Mehta ,Varsha Marathe ,Priyanka Padvi and Manjusha Shewale

MITCOE,,Savitribai Phule Pune University,Pune

Maharashtra,India

Abstract: **With the increase in amout of data and information one has to deal with,now a days,going through all the documents is a time consuming process.We are implementing an android application that helps organizations such as law firms to manage the hundreds of documents and to get summary of these documents.We are also using concept of ontology for this application.Ontology is basically the relationship between entities.The application that we are implementing allow the users to search for files in the database,upload files and summarize multiple documents.**

Keywords: Term Frequency,Inverse Document frequency,stop words,stemming,android ,multiple documents, summarization.

1.INTRODUCTION

Now a days, with the enhancement in data mining and information retrieval systems,the amount of data one has to deal with has increased rapidly. The number of documents to take into consideration are far more and the time to go through all such documents is very less. If the important information present in these documents are made available in condensed form,the user saves time and the user can focus on important parts of the documents easily. We have implemented an application,that can be used in organizations for this very purpose. All the users have to register with the server to access the functions provided by the server.

Now, any of these users can upload useful documents on the server which they would like to go through.

- When a user uploads a document, all the other users receive a notification that a new document has been uploaded.
- User can search for a document he wants to go through.
- If the user is short of time,user can get summary of the documents he wishes to read.

When a user uploads a document, all the other users receive a notification that a new document has been uploaded.

User can search for a document he wants to go through.

If the user is short of time,user can get summary of the documents he wishes to read.

II. RELATED WORK

A number of research study have addressed multi document summarization.Many approaches and systems are available for multi document summarization.Given below is some research study from related literature about these approaches and systems for multi-document summarization.

A. Feature based method:

Extractive type summarization includes identifying the relevant sentences from the text and put them together to create an accurate summary. Features ithat influence the importance of sentences are determined .Some of the features that are considered for selection of sentences are: Title or headline word ,location of sentence, length of sentence etc.

B.Word frequency:

The basic idea of using word frequency is that important words are found many times in the document. Tf and idf is one of the most common measure used to calculate the word frequency.

C.Cluster based method:

Basically,clustering is to group similar objects into their classes.For clustering of multi documents, these objects refer to sentences and the cluster that a sentence belongs to is represented by classes.

The cosine similarity measure is one of the common techniques used to measure similarity between a pair of sentences. Here sentences are represented as a weighted vector

D.Graph based method:

Agraph can be represented in the form of G = (V, E), here V is vertex or node in the graph and E is the edge between each vertex. In case of text documents, vertex represents sentence and edge is the weight between two sentences. Using this approach, documents can be represented as a graph where each sentence becomes the vertex and the weight between each vertex is the similarity between the two sentences.

Given below are some of the systems which are used for multi-document summarization.

E.Newsblaster

Newblaster is a system that helps users find news that the user may find interesting.Newblaster automatically does the collection, clustering, categorizing and summarizing of news from many websites on a daily basis, and also it provides interface for the users to browse the results.

F.iResearch Reporter

iResearch Reporter accepts a query from the user and passes this query to Google search engine, multiple relevant documents are retrieved, categorizing of these documents is done,and readable natural language summary reports are produced that cover multiple documents in the retrieved set.

G.Ultimate Research Assistant

Text mining on Internet search results is performed for summarizing and organizing them and hence performing online research becomes easier for the user.

III. ARCHITECTURE:

Given below is the architecture of the system we are implementing:

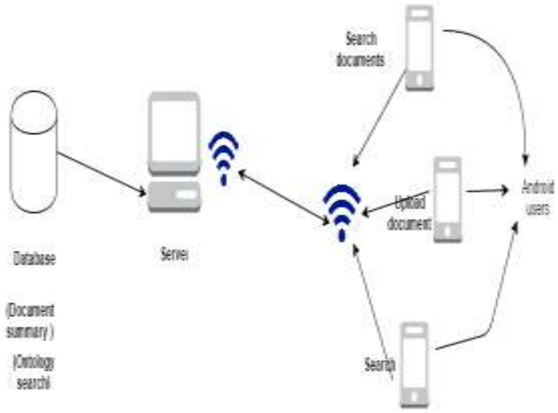

Fig.1 Architecture of system

IV.IMPLEMENTATION

A.UPLOAD AND SEARCH:

First of all ,to carry out the functions of upload and search ,we have implemented a web server using java.Basically a web server is a software which is responsible for accepting client requests,retrieving the specified file and returning its contents.A Java servlet is a Java program that extends the capabilities of a server.Java servlets are used to enable a user to upload and search documents. A Servlet can be used with an HTML form tag to allow users to upload files to the server and fetch files from the server. An uploaded file could be a text file or image file or any document.

B.STOP WORDS REMOVAL:

In data mining applications, we often come across the term “stop words” or "stop word list" or even "stop list". Stop words are basically a set of commonly used words, not just in English but in any language Stop words are important for many applications because, if the words that are very commonly used in a language are removed,it helps us to focus on the important words. For Stop word removal we are using pattern matching algorithm.

- Pattern matching algorithm:

Pattern Matching algorithm can be explained with the pseudo code given below:

```
//Pseudo Code: do

if (text letter == pattern letter)

compare next letter of pattern to next letter of text else move pattern down text by one letter while (entire pattern found or end of text)
```

C. STEMMING:

In data retrieval process basically, stemming is the process of reducing inflected words to their word stem, base or root word.

For example:if the term is engineering the stem will be engineer.Also for term engineered,the stem will be engineer.

To improve recall by automatic handling of word endings is the main aim of stemming. At the time of indexing and searching ,this is done by reducing the words to their word root.

For example, if a user includes the term” *stemming”* as part of a query, it is possible that the may also be interested in variants of stemming such as *stemmed* and *stem*.

Stemming plays an important role in improving the efficiency of the application.

For stemming,we are using dictionary stemmers.

In Algorithmic stemmers, a standard set of rules is applied to each word,whereas in dictionary stemmers,the word is simply looked up for in the dictionary. Theoretically, dictionary stemmers produce much better results than an algorithmic stemmer. Following are the main functions of a dictionary stemmer:

- Even for words such as feet and mice it should return the correct root word.

- Recognizing the difference between words that are similar but have different meanings—for example, organ and organization

The meanings of words change with time. While stemming mobility to mobil may have would have made sense some years back,but now it conflates the idea of mobility with a mobile phone.

Thderefore it is important to keep the dictionaries current.This could be a time-consuming task.Many a times, by the time a dictionary is made available, some of its entries are already out-dated.

The quality of the dictionary is therefore decided by this.

The process of removing prefixes and suffixes may be more or less efficient,depending on the quality of the dictionary.

D.SUMMARIZATION:

Now summarization of documents is the most important part of our application.

For summarization of documents we are implementing TF-IDF algorithm.

TF-IDF Algorithm :

Term frequency–inverse document frequency, is basically a numerical statistic that is meant to show how important a word is to a document in a collection of documents.In information retrieval systems,it is used as a weighting factor. As the number of times a word appears in the document increases,the tf-idf value increases proportionally. But this tf-idf value is decreased by the frequency of the word in the collection.This helps to take into account the fact that some words appear more frequently in general.

For the value of term frequency tf(t,d), the most easy way to go is to use the *frequency* of a term in a document, i.e. the number of times that term t repeats in document d. If $f_{t,d}$, denotes the raw frequency of t then the simple tf scheme can be given as tf(t,d) = $f_{t,d.}$

The inverse document frequency basically measures the amount of information provided by a word, that is, whether the term is common or rare across all documents. It is a logarithmically obtained value. The total number of documents in the collection is divided by the number of documents containing the term.The log of this term is then calculated.The value obtained is idf.

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Then tf–idf is calculated as:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

E.USE OF ONTOLOGY

In computer science and information science, an ontology is a formal naming and definition of the interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.In our application,Ontology is used to define the relationship between the search query and the high frequency terms in the database.

For example:

If the search query is “Organization” Ontology defines the relationship between this search query and the similar meaning words in our database.Similar words present in our database can be:“Institution”“Corporation” etc. For this purpose we use a tool called Wordnet. WordNet is a lexical database for the English language. It groups English words into sets of synonyms called *synsets*, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.its primary use is in automatic text analysis and artificial applications.

V.APPLICATIONS:

Can be used in any organization where loads of documents are to be handled.

Can be used in Law Firms.

Can be used in Hospitals for in house patients.

VI.FUTURE SCOPE

The application can be improved to provide a more accurate summary.

The application can be improved to increase the amount of data and number of documents it can handle.

It can be extended to make compatible with windows , Symbian , Bada etc.

VII.CONCLUSION

TF-IDF algorithm has been executed to determine frequency of words.

Thus, we have used word-frequency based approach for multi document summarization.

An android application has been implemented to access the functionalities provided by the project.

VIII.REFERENCES

1. H. Hsu, C. Tsai, M. Chiang, and C. Yang, “Topic generation for web document summarization,” inProc. IEEE SMC, 2008, pp. 3702–3707.
2. L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S. Chen, and V. Hristidis, “Using data mining

techniques to address critical information exchange needs in disaster affected public-private networks,” in Proc. SIGKDD, 2010, pp. 125–134.

3. An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management Lei Li and Tao Li-IEEE transactions on systems, man, and cybernetics: systems, vol. 44, no. 2, february 2014.

IX.AUTHOR PROFILES

Amit Savyanavar -Professor in Computer Engineering Department of MIT College of Engineering, Savitribai Phule Pune University, Pune Maharastra,India.

amit.savyanavar@mitcoe.edu.in

Bhakti Mehta- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.

bhaktimehta0909@gmail.com

Varsha Marathe- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.

varshamarathe7979@gmail.com

Priyanka Padvi- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.

priyu.padvi@gmail.com

Manjusha Shewale- B.E. in Computer Engineering from MIT College of Engineering,Savitribai Phule Pune University,Pune,Maharashtra,India.

manjusha.shewalen24@gmail.com