MULTIKEYWORD HUNT ON PROGRESSIVE GRAPHS

Iswarya.S[1], Pavithra.M[2], Selvakrishnan.S[3]

[1,2] Jeppiaar SRR Engineering College.
[3]Selvakrishnan.S, Dept of IT, Jeppiaar SRR Engineering College, Tamil Nadu, India.

---***---

Abstract- *Data mining or Knowledge Discovery is the process of analyzing data from different perspectives and summarizing it into useful information. This information can then be used to increase revenue, cuts costs, or both. A software created with Data mining as its basic theme should allow users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information and for content providers to classify and catalog documents. In order to deal with these problems, a look towards automated methods of working with web documents so that they can be more easily browsed, organized, and catalogued with minimal human intervention. In this project, visualization technique is used to visualize topically similar documents.*

KeyWords: *Top K-query, Continuous query; Document stream.*

1. INTRODUCTION

Traditional web search engines often return hundreds or thousands of results for a search, which is time consuming for users to browse. On-line libraries, search engines, and other large document repositories *(e.g.* customer support databases, product specification databases, press release archives, news story archives, *etc.)* are growing so rapidly that it is difficult and costly to categorize every document manually. In order to deal with these problems, a look towards automated methods of working with web documents so that they can be more easily browsed, organized, and cataloged with minimal human intervention. In contrast to the highly structured tabular data upon which most machine learning methods are expected to operate, web and text documents are semi-structured. Web documents have well-defined structures such as letters, words, sentences, paragraphs, sections, punctuation marks, HTML tags, and so forth. It is estimated that as much as 85% of all digital business information, most of it web-related, is stored in non-structured formats *(i e .*non-tabular formats, such as those that are used in databases and spreadsheets *)*. Developing improved methods of performing machine learning techniques on this vast amount of non-tabular, semi-structured web data is therefore highly desirable. Clustering and classification have been useful and active areas of machine learning research that promise to help us cope with the problem of *Graph-Theoretic Techniques for Web Content Mining* information overload on the Internet. With clustering the goal is to separate a given group of data items (the data set) into groups called clusters such that items in the same cluster are similar to each other and dissimilar to the items in other clusters.

2. PROBLEM IDENTIFIED

In the present scenario, keyword-based search mechanisms are used to find a particular topic. The drawback in this technique is that, one can identify relevant information only when there is some basic knowledge on the topic. It is impossible to learn about a particular topic, without any relevant keywords to obtain information. In order to overcome this situation, visualization technique is used to visualize topically similar documents. The topics will be displayed as a cluster and the topics are differentiated based on colours and positioned in an attempt to keep the related topics close to each other. The documents that are semantically similar to a particular topic is found. These topics are presented in the form of a visualized graph and it indicates the degree of closeness between relative topics. Iterative search mechanisms are implemented until the desired result is obtained.

3. EXISTING SYSTEM

A general approach to protect the data confidentiality is not there so information's may leak. Existing system uses key word based search. Summarization techniques are generally sentence based or key word based. It does not directly provide information about relationship between documents. Computations are memory intensive and the resulting topics are not easily interpreting table. In this existing system graph visualization applications require a large amount of processing power. Traditionally, use of document clustering was viewed as an inefficient way to search through large corpuses because of complexity issues. In existing a user can able to search with the help of a filename only. Huge cost in terms of data usability. For example, the existing techniques on filename-based information retrieval. Finding relationship between the document and document itself is not confidentiality Possible.

4. PROPOSED MODEL

In our Proposed System we are using an algorithm called STEMMING which is used for Recommending that particular file for users. In Proposed we are using a Pre-processing Technique which is used to Remove the Stop words from the file and finding the Root Words of it. Using Re-Ranking concept the frequent words which is present in the particular document will be listed as Top 5 and shown in a Crystal Report. To provide better security, here in our application we are generating an OTP (One Time Password) and that OTP was sent to registered mail-id for secure login. Fast Retrieval of documents over existing schemes. Attacks and Data losses are avoided due to providing security. We using an advanced encryption scheme that supports both the accurate multi-keyword ranked search and flexible on dynamic operation on document collection.

5. PROPOSED DIAGRAM

The proposed model block diagram is represented as follows:

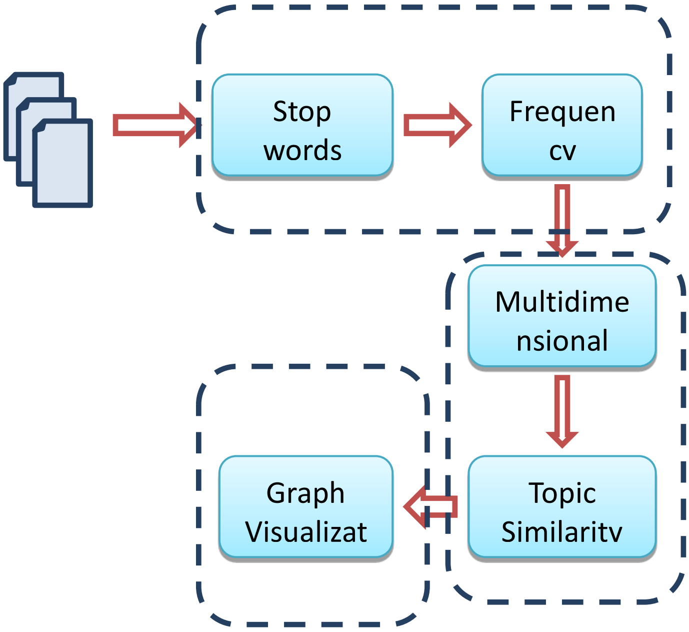


Fig: Architecture Diagram

REQUIREMENT SPECIFICATION

5.1 HARDWARE REQUIREMENTS

- SYSTEM : Pentium IV 2.4 GHz
- HARD DISK : 500 GB
- RAM : 4 GB

5.2 SOFTWARE REQUIREMENTS

- Operating system :Windows
- IDE :Microsoft Visual Studio 2012
- Database : SQL server 2014
- Coding Language: ASP C#.NET.

5.3.1LOAD DATASET

A dataset (or data set) is a collection of data. The dataset may comprise data for one or more members. For example, 20 Newsgroups and Reuters21578 top-10, which are two real-world data sets, both have more than 15,000 features. This project uses 20_Newsgroups data set from UCI repository. The data sets in this repository are of variouscategories,"alt.atheism","comp.graphics","comp.os.mswindows.misc","comp.sys.ibm.pc.hardware","comp.sys.mac.hardware","comp.windows.x","misc.forsale","rec.autos","rec.motorcycles","rec.sport.baseball","rec.sport.hockey","sci.crypt","sci.electronics","sci.med","sci.space","soc.religion.christian",etc, are the various news datasets. This system uses "sci.electronics","sci.med" and "sci.space" datasets.

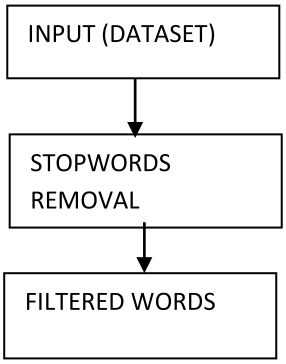
5.3.2 DATA PREPROCESSING

5.3.2.1STOP WORDS REMOVAL

Stop words are words which are filtered out prior to, or after, processing of natural language data (text). It is controlled by human input and not automated. These are some of the most common, short function words, such as the, is, at, which and on. Stop Words can cause problems because they increase redundancy in the given document to be processed. Therefore removal of these stop words is necessary in a Question Answering System to make the retrieval process less complex.

This project uses nearly 600 stop words, to remove unnecessary words from the dataset that has been loaded.

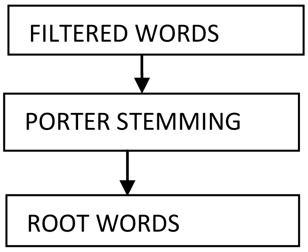
BLOCK DIAGRAM



5.3.2.2 PORTER STEMMING

Stemmers employ a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned. Eg: A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

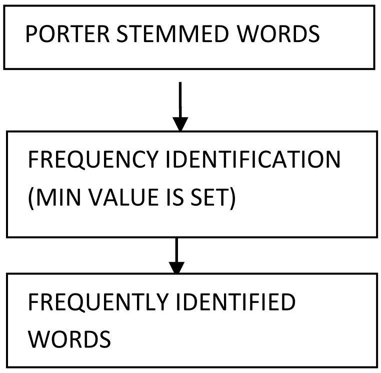
BLOCK DIAGRAM



5.3.3FREQUENCY IDENTIFICATION

This project combines the definitions of term frequency and inverse document frequency to produce a composite weight for each term in each document. As the traditional dimensionality reduction methods for text clustering, such as document frequency, mutual information, information gain. The word count is found and the minimum value is set, from this the most frequently used words are obtained.

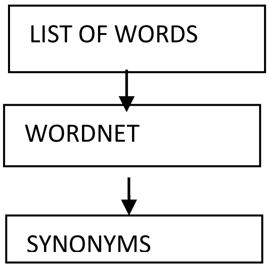
BLOCK DIAGRAM


PORTER STEMMED WORDS
FREQUENCY IDENTIFICATION (MIN VALUE IS SET)
FREQUENTLY IDENTIFIED WORDS

5.3.4 SYNONYMS GENERATION

Synonyms Generation is a technique which is used to find out the similarity in data. Synonyms are words that are similar or have a related meaning to another word. Words that are synonyms are said to be synonymous, and the state of being a synonym is called synonymy. Word Net tool is used in this system to generate synonyms of the most frequently used words.

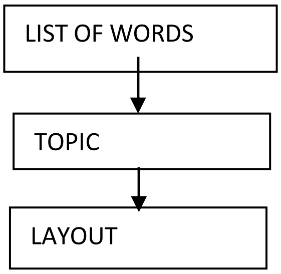
BLOCK DIAGRAM



5.3.5 FORCE DIRECTED LAYOUT

Force directed layout method is used to position the topics according to the relationship between them. When synonyms are generated for list of words, the words with same meanings are grouped and each group is given a common name which is known as Topic. These Topics can still be grouped and placed under another Topic.

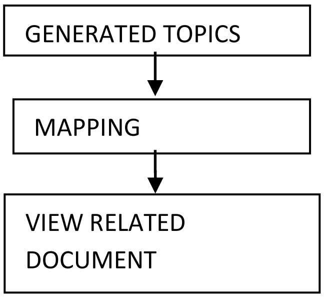
BLOCK DIAGRAM



5.3.6 GRAPH VISUALIZATION

In Graph Visualization technique the document related to a topic is displayed i.e. a list of main Topics will be displayed. When a particular Topic is selected, sub-topics related to that Topic will be displayed. Further, when a sub-topic is selected, a document related to that sub-topic is displayed.

BLOCK DIAGRAM



6. CONCLUSION

The visualization technique has been developed for large set of documents using statistical topic model. This system requires dataset that has to be loaded, and it works only on large sets of data but not on queries. The graph visualization mechanism can also be incorporated using several techniques such as topic-based deformation of ordered sets of nodes in a graph, collapsing of nodes based on semantic association, single or multiple topic influences, interaction with topic and document nodes through interpolation over mouse movements, iterative text-search-and-visualization steps, cluster based on topic similarity, and a range of more generic graph interaction methods.

REFERENCES

1. Asuncion, H., Asuncion, A., and Taylor, R.Software traceability with topic modeling. In Proceedings of the 32nd ICSE Conference. ACM, 2010.

2. Cao, N., Sun, J., Lin, Y.-R., Gotz, D., Liu, S., and Qu, H."Facetatlas: Multifaceted visualization for rich text corpora". IEEE Trans. Vis. Comput. Graph. 16, 1172–1181, 2010.

3. Cutting, D. R., Karger, D. R., and Pedersen, J. O. "Constant interaction-time scatter/ gather browsing of very large document collections". In Proceedings of the 16th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR'93). ACM, New York,126–134,1993.

4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R."Indexing by latent semantic analysis".J. Amer. Soc. Inf. Sci. 41, 391–407,1990.

5. Eick, S. G. and Wills, G. J."Navigating large networks with hierarchies. In Proceedings of the 4thConference on Visualization (VIS'93)". IEEE Computer Society, 204–209, 1993.