www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 5 May 2015, Page No. 11926-11931

Optimization of Search Engine contents

Gaurish Kamat [1], Viraj Kochure [2], Tushar Patil [3], Utkirna Shinde [4], Leena Deshpande[5]

Vishwakarma Institute of Information Technology, Pune

[1]gaurishkamat23@gmail.com

[2]virajkochure@gmail.com

[3]patiltushar9997@gmail.com

[4]utkirnas05@gmail.com

[5]deshpande.leena27@gmail.com

Abstract –

As an information resource the Internet is increasing in size, depth and complexity. Availability of information is no longer an issue on the Internet. In an era where human presence is worth so much, it is becoming increasingly time consuming to analyze the relevant information from an ever-growing sea of inadequate or inappropriate data and accelerate the laborious process of web browsing. Information provided by search engines to web users is irrelevant, due to the lack of providing structural information and categorization of the documents. The issue of uncategorized data has become a standoff due to the inconsistencies and variations in the characteristics of the data.

In this paper, we present a way to cut short the sea of data to relevant data by forming categories and then providing summarized data with links in specific category. The projected way has insightful ability to improve the drawback and returns efficient outcomes.

Index Terms *- Relevant data, k-means, categorization.*

1. Introduction

Search Engine is software that searches for data based on some criteria. Every Web search engine site uses a search engine that it has either developed itself or has purchased from a third party. Search engines can differ dramatically in the way they find and index the material on the Web, and the way they search the indexes from the user's query. Although a search engine is technically the software and use algorithms to perform a search. For example, Google [12] is a major search site on the Web, but rather than being called the "Google search site," it is commonly known as the "Google search engine."

Since we live in a computer era, Internet has become a part of our everyday lives and information is only a click away. Just open your favourite search engine, like Google, Bing, Yahoo, type in the key words, and the search engine will display the pages relevant to your search. During the past two decade the frequency at which technology boosted is very high. From the first 3 web sites that the baby Internet had (Microsoft, Netscape, and Amazon), to 800 million pages in 1999 and to over 50 billion pages today, the Internet has experienced an exponential growth. Even the simple surf on the Internet is enough to convince you that there is a huge amount of information and links available online. However, all this information is useless unless we have a way of searching and sorting it.

1.1 Earlier Scenario

From the first search engine Archie in 1990, to the modern search engines we use today, the problem of

deciding the relevancy of the information available online has been an issue. This used to be the correct picture in the early 90s, when the first search engines used *text based ranking systems* to decide which pages are most relevant to a given query. There were however a number of problems with this approach. Suppose we wanted to find some information about Apple. We type in the word "Apple" and expect that "www.apple.com" would be the most relevant site to our query. However there may be millions of pages on the web using the world Apple, and www.apple.com may not be the one that uses it most often. Suppose we decided to write a website that contains the word "Apple" a billion times and nothing else. Would it then make sense for our web site to be the first one displayed by a search engine? The answer is obviously no. However, if in all search engine does is to count occurrences of the words given in the query, this is exactly what might happen.

1.2 Recent Scenario

Now-a-days search engines use multiple algorithms for efficient search. For example, Google uses PageRank Algorithm [12], Knowledge Graph [11] to give information. The only issue with this is it gives information randomly. Like when we search it retrieves information from almost every field irrespective of what is required by the user. It becomes very hectic for the users' to go through each and every link and analyse that is it according to what is required. Suppose we wanted to find information about Synchronisation. Google [12] will provide number of links with different content like synchronisation meaning, data synchronisation, synchronisation in java, synchronisation of alternators, synchronisation in operating system, synchronisation of generators, process synchronisation etc. If user wants information on synchronization in java, he will have to open different links and go through whether it is according to what is required or not. It would be a hectic job to go through each and every link and check the content.

Instead providing categories with summary and links will cut short the hectic job and searching time and will increase efficiency. It will play a vital role in many information management and retrieval tasks. It can also help to improve the quality of web search.

2. RELATED WORK

People who rely on internet for information and data frequently turn to "Google"; almost a synonym to Internet to many. For last 15 years Google has become the most widely used search engine [5]. At first Google could only retrieve websites. It was only after a little flourishing, that they were able to implement various supporting aspects like news, video search, Image search etc.

As the Google average per day search has increased from 9800 to 5,134,000,000 the amount of energy required to run search engine has also increased. Every search cost around 0.2-7 grams of carbon emission. To reduce this emission Google is working on this to ensure responsible use of Google [13].

The idea of splitting a Web page to fragments has been used by Cai et al. [9], Lee et al. [15] and Song et al. [16], where they extract query-independent rankings for the fragments, for the purpose of improving the performance of web search and also to facilitate web mining and accessibility. Cai et al. [9] partition a web page into blocks using the vision-based page segmentation algorithm. Based on block-level link analysis, they proposed two new algorithms, Block Level PageRank and Block Level HITS to extract authoritative parts of a page. Lee et al. [15] discuss a Web block classification algorithm after Web page division into semantic blocks, while Song et al.[4] provide learning algorithms for block importance.

A small set of rules were manually constructed by observing a limited set of blogs from the Blogger and Word press hosting platforms. These rules operate on the DOM tree of an HTML page, as constructed by a popular browser Mozilla Firefox [3] Chekuri et al. studied automatic web page classification in order to increase the precision of web search. A statistical classifier, trained on existing web directories, is applied to new web pages and produces an ordered list of categories in which the web page could be placed. here the classification category have been studied in the different point of view like binary, multiclass, flat, hierarchical, soft and hard classification.

3. PROPOSED SOLUTION

In this research paper a framework is proposed that gets text input from the user in the form of single or multiple keywords. Input is analyzed and processed to retrieve information. The retrieved information is then categorized and displayed to the user. Now information is just a click-away. The user can click on any of the category to get the desired information. The proposed architecture not only provides categories but also displays the summarized document of the category selected.

The purpose of Search Engine optimization is to allow effective search on the collection of Unstructured Text. It might be helpful to many users. It will help people who are not used to Internet and old age people who are not aware of using the internet for search purpose. It will also reduce the task of going through each and every link and gathering the data. Thus it will help to improve the quality of web search.

4. ARCHITECTURE OF THE SYSTEM

In this research paper a framework is proposed that gets text input from the user in the form of single or multiple keywords, understands the input and pass it to the third party. Third party can be any search engine like Google, Yahoo, and Bing etc. Third party already has its indexed database and algorithms to retrieve information. It retrieves information in form of links and html data. That data is downloaded in JSON format and then links are extracted using JSON parser. Links are crawled and all the web documents are converted from HTML files to normal plain text file. After removing the HTML tags and the textual data is then stored in text files which is done with the help of HTML parsers. The textual part can be found in any of these tags, thus extracting the necessary information from tags like .. , <pre>..</pre> etc. will be done. All text files are merged into a single text file. This text file can vary in size depending on the retrieved information from third party. Now, the obtained text file is pre processed and Term Frequency (TF) and Inverse Document Frequency (IDF), cosine index is calculated and passed to k-means for clustering.

Clustering is unsupervised classification of data set to reduce the amount of data by categorizing or grouping similar data items together. To apply the clustering techniques, each document is usually represented as a vector of weighted term frequencies such as Term frequency (TF) and Inverse Document Frequency (IDF). For these vectors, it is necessary to calculate a similarity or distance measure that clustering algorithm defines between two vectors. For these calculations, a pair of closest points is merged into a new single cluster. This merge process is repeated until a stopping criterion is satisfied.

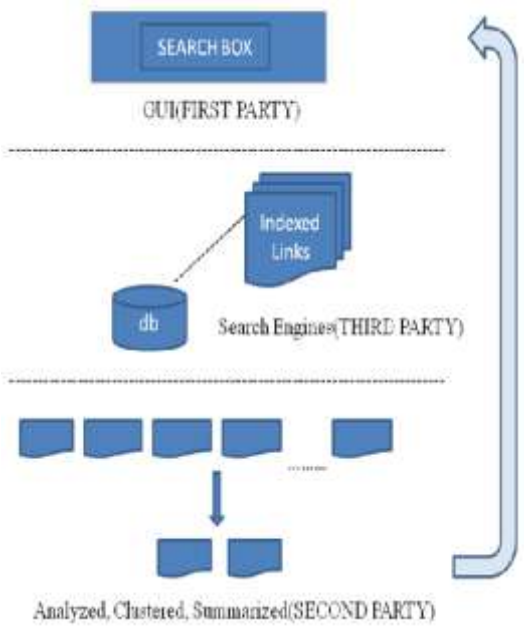

Figure 1: Basic Architecture of the System.

5. SYSTEM'S WORKFLOW

The workflow of system consist of different steps that are connectivity to search engine through API, Extract text data from links, Preprocessing of data, forming categories from data and provide summary with links. Here is the detailed description of the workflow.

The search results retrieved from the third party are in the form of links (URLs). These URLs are visited separately and the data is extracted using the HTML parsers. Data from all the URLs is gathered into a single document. After combining data into a single file the file is preprocessed and stemming is done by using porter's algorithm. After filtering the term frequency is calculated for each and every word in the file. Term frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is often used as a weighting factor in information retrieval and text mining. Then the file is evaluated by calculating the similarity measures like cosine and dice which helps us to identify important and unique sentences from a file to form various clusters. Then the resorting of clusters sentences is done to form the clusters with relevant data. Then the summary of that data will be generated for each cluster. The diagram below depicts the overall flow.

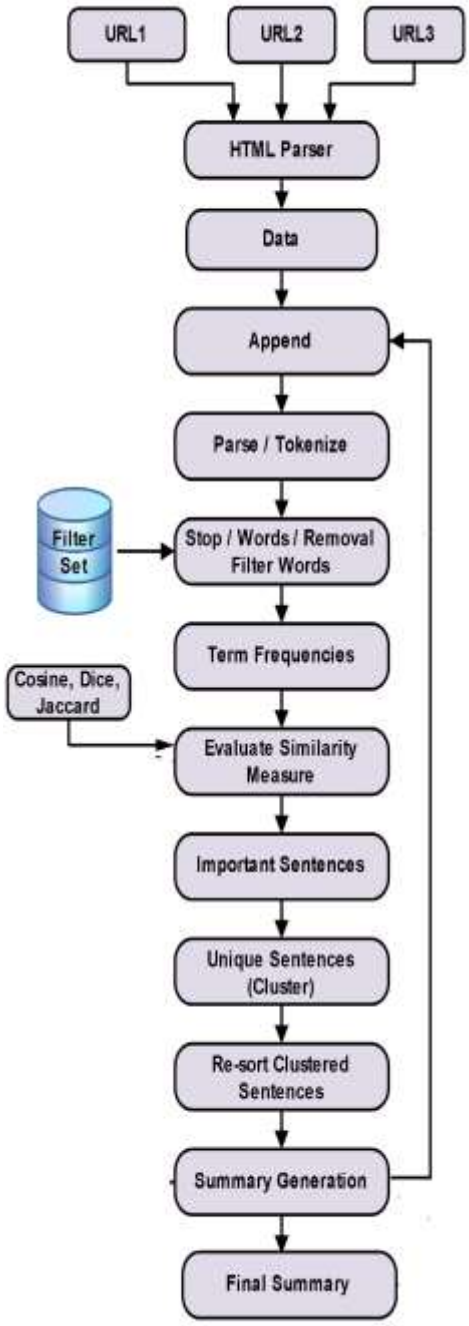

Figure 2: System Workflow

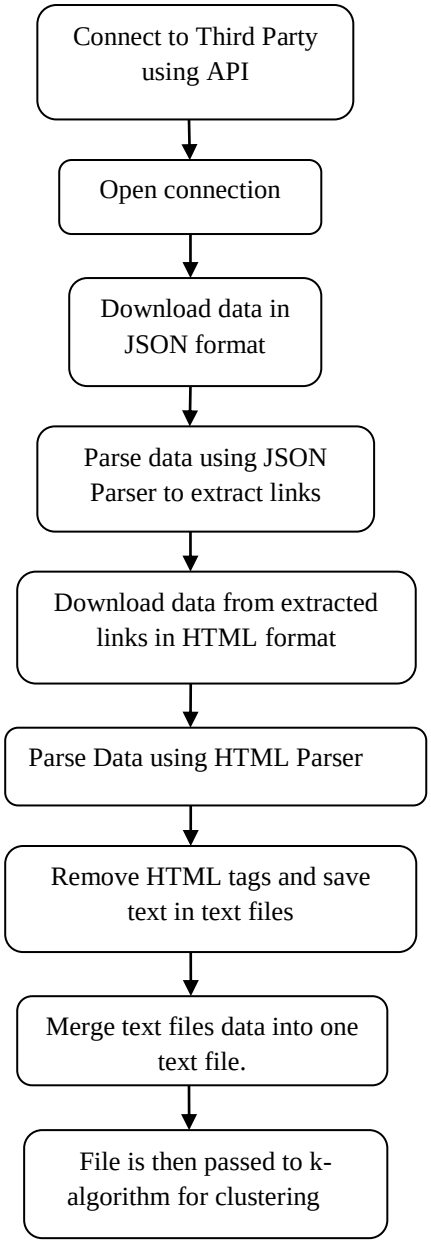

Figure 3: Data Extraction

4.1 Data extraction from Third Party

Web search engines work by storing information about a large number of web pages, which they retrieve from the WWW itself. These pages are retrieved by a web crawler an automated web browser which follows every link it sees, exclusions can be made by the use of robots.txt. Data about web pages is stored in an index database for use in later queries.

We retrieve information of third party i.e. search engine by using an API (Google Custom Search API). We download the data from the links and perform parsing to store it into the text files.

4.2 Page categorization algorithm building blocks

k-means[6] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters).

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

4.3 OUTPUT

We have used software tools like JAVA, J Frames and NETBEANS IDE in our work for information retrieval and page categorization.

4.3.1 Data Retrieval

Data Retrieval is based on the search engine content for top 10 URLs to identify the correctness of Data Retrieval.

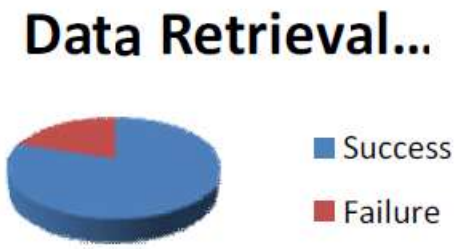

Figure 4: Data Retrieval

Data retrieved from API is slightly different from what Search Engine retrieves. Correctness of data retrieved from Search Engine is more than what API retrieves. Also amount of data differs. For example if search engine retrieves 200 keywords then keywords retrieved by API will be less than what is retrieved by Search Engine.

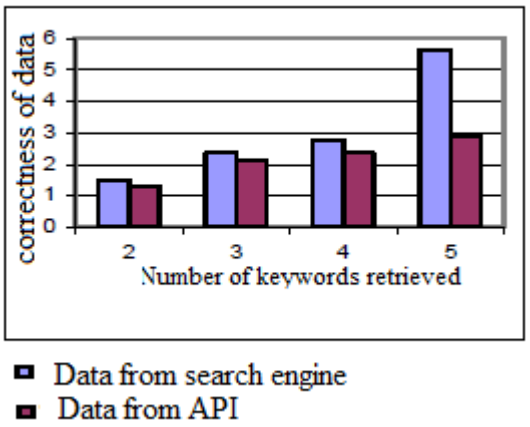

Figure 5: Correctness of data retrieved

4.3.2 Clustering of data points

Clusters are based on the test data for top 10 selected URLs to identify the correctness of algorithm.

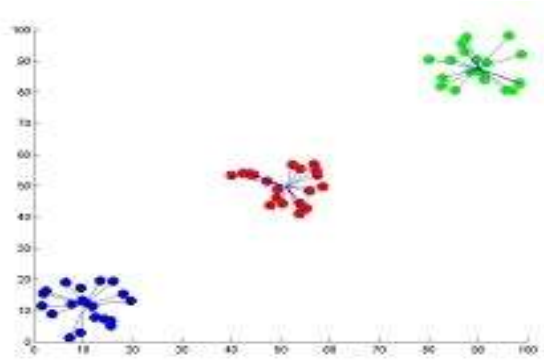

Figure 6: Showing the result of k-means for *'N'* = 60 and *'c'* = 3

5. CONCLUSION & FUTURE WORK

The designed algorithm is used to understand the user's information and extract the relative information. It has a perceptive ability to understand the user requirements and then search and extract the refined results in form of categories and to present query-specific summaries for text documents.

Improvement in the algorithm can improve the accuracy ratio of the result and can ultimately influence the accuracy ratio of searching results. In the future, we plan to extend our work to account for links between documents of the dataset. For example, exploit hyperlinks in providing summarization on the web.

References:

[1] J. Pasternack and D. Roth. Extracting article text from the web with maximum subsequence segmentation. In WWW '09: Proceedings of the 18th international conference on World wide web, page 1971{980, New York, NY}
[2] Yang Mingqiang, KpalmaKidiyo and Ronsin Joseph (2011),"Chord Context Algorithm for Shape Feature Extraction", Object Recognition, Dr. Tam Phuong Cao (Ed.), ISBN: 978-953-307-222-7.
[3] Data Extraction and Web page Categorization using Text Mining, International Journal of Application or Innovation in Engineering & Management (IJAIEM) ISSN 2319 – 4847, Volume 2, Issue 6, June 2013, Department of Computer science & Engineering.
[4] Xiaoguang Qi and Brian D. Davison Department of Computer Science & Engineering Lehigh University Web Page Classification: Features and Algorithms June 2007.
[5] Baroni, M., Chantree, F., Kilgarri, A., Sharo, S. (2008). Cleaneval: a competition for cleaning web pages. In Proceedings of the 4th Web as Corpus Workshop (WAC4), - Can we beat Google?. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, Proceedings of the 6th International Language Resources and Evaluation (LREC 2008). Marrakech, Morocco, 2008
[6] Web Page Classification Using Data Mining, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, ISSN (Print) : 2319-5940 ISSN (Online) : 2278-1021, Issue 7, July 2013
[7] Qasem A. Al-Radaideh, Eman Al Nagi "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 3, No. 2, 2012
[8] Pattern Classification based on Web Usage Mining using Neural Network Technique, International Journal of Computer Applications (0975 – 8887) Volume 71– No.21, June 2013.
[9] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A System For Keyword-Based Search Over Relational Databases. ICDE, 2002.
[10] M. Asif Naeem, Noreen Asif,A Web Smart Space Framework for Intelligent, International Journal of Emerging Sciences ISSN: 2222-4254 1(1) April 2011,Search Engines,Department of Computer Science,University of Auckland,NewZealand.
[11] Google's Knowledge Graph: Key Take Aways from the New Search Functionality, Written by Angel Sancho Ferrer on May 29, 2012.
[12] Kurt Bryan, Tanya Leise, The $25,000,000,000 Eigenvector: The Linear Algebra behind Google, SIAM Review, Vol. 48, No. 3. (2006).http://www.siam.org/journals/sirev/48-3/62328.html
[13] Effective use of Google. (CSI Journal August 2013).
[14] Text mining Ian H. Witten Computer Science, University of Waikato, Hamilton, New Zealand email ihw@cs.waikato.ac.nz.
[15] C.H. Lee, M.Y. Kan, S. Lai: Stylistic and Lexical Cotraining for Web Block Classification. WIDM, 2004
[16] R. Song, H. Liu, J. Wen, W. Ma: Learning Block Importance Models for Web Pages. WWW, 2004