Implementation of Mini-Search Engine

Khushboo Pandey[1] Priyanka Khanka[2] Sakriti Karan[3] Neha Kapadia[4]
[1,2,3,4]Department of Information & Technology (IT)
[1,2,3,4] Thakur College of Engineering and Technology Mumbai – 400101, India

Abstract----Search Engine can be defined as a program that searches for and identifies items in a database that correspond to keywords or characters specified by the user, used especially for finding particular sites on the Internet. Search engines retrieve information using algorithms such as distance vector algorithm, crawlers, meta-tags, indexing and many such others based on the keywords or queries entered by the user. When the user queries a search engine to locate information, he/she is actually searching through the index that the search engine has created — not actually searching the Web. These indices are giant databases of information that is collected and stored and subsequently searched. This is why sometimes a search on a commercial search engine, such as Yahoo! or Google, returns results that are, in fact, dead links. Since the search results are based on the index, if the index hasn't been updated since a Web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated. The overall goal of this project is to develop a scalable, high performance search engine. The main focus is on the algorithmic challenges in compactly representing a large data-set while supporting fast searches on it. Our intention is to cluster different documents based on subjective similarities and dissimilarities. Our proposed tool 'Mini Search Engine' is based on the concept of data mining, page ranking algorithm and word search program. It presents results in different file formats like .pdf, .doc etc. based on the user's query.

Keywords: data mining, page ranking, queries

I. Introduction

Search engines are the key to finding specific information on the vast expanse of the World Wide Web. Without sophisticated search engines, it would be virtually impossible to locate anything on the Web without knowing a specific URL. We intend to create a simple yet very powerful and fast search engine. The following section is a detailed description of the project consisting of the following:

- Data Files: A description of the format of the data files used for the search engine. The input files consist of text documents, images, videos.
- The Program: In addition to the input files an initial and very simple search engine will be programmed.
- GUI Design and implement a graphical user interface for the search engine.
- Space Efficiency Improve the space usage of the data structures. Improving the space will also have a significant effect on the preprocessing and query time due to effects of the memory hierarchy.
- Compression Further improve the space usage by using advanced compression techniques.
- Boolean Search Implement Boolean searches with operators such as. For instance, "latex AND typesetting" should find all documents containing both latex and typesetting.
- Ranking Report the result documents ordered by the rank of the different documents. The rank of a document could be the determined from the number of occurrences of the search terms in the document.
- Auto-completion given the prefix of a search string suggest possible endings automatically. Particularly relevant when combined with a GUI.
- Crawler Write a program to automatically extract information from some source, e.g., the internet, into suitable data files for a search engine.
- Java Data Structures Implement the data structures for the index using the standard Java collection classes.
- Spelling Suggestion If a search does not match any search string it may be because the user misspelled the search string. Construct a mechanism to suggest alternatives that almost match the search query.

II. Page Ranking Algorithm

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google:

PageRank works by counting the number and quality of link to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known. Google uses an automated web spider called Googlebot to actually count links and gather other information on web pages. PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by $PR(E)$ other factors like Author Rank can contribute to the importance of an entity.

A PageRank results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

Numerous academic papers concerning PageRank have been published since Page and Brin's original paper. [4] In practice, the PageRank concept may be vulnerable to manipulation. Research has been conducted into identifying falsely influenced PageRank rankings. The goal is to find an effective means of ignoring links from documents with falsely influenced PageRank. [5]

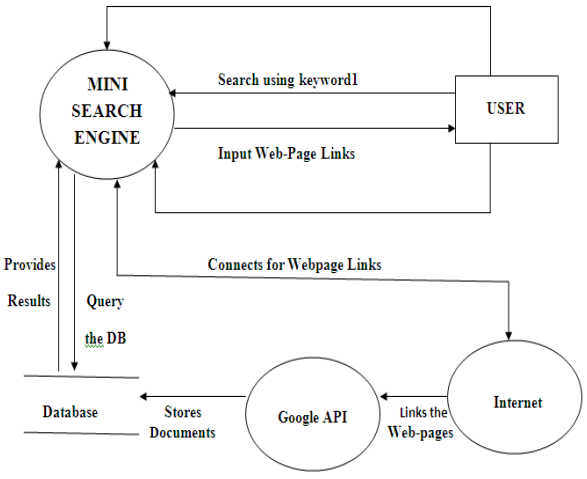
Other link-based ranking algorithms for Web pages include the HITS algorithm invented by Jon Kleinberg (used by Teoma and nowAsk.com), [citation needed] the IBM CLEVER project, the TrustRank algorithm and the hummingbird algorithm

III. Data Mining

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

IV. Implementation



DFD LEVEL 0

V. Future Work

In the future, we will build our database which includes data within and direct search would be made instead of just giving the links.

Our system can also perform the following searches:

- Image
- Audio
- Video
- Privacy: When you search Google, and click on a link, your search term is usually sent to that site, along with your browser & computer information, which can often uniquely identify you. Those sites usually have third-party ads, and those third-parties build profiles about you. That's why those ads follow you everywhere. Your profile can also be sold and potentially show up in unwanted places like getting insurance. Google also saves these searches. Your saved searches can be legally requested and then used against you. Also a bad Google employee or hacker could go snooping.Whereas our search engine does not send your searches to other sites or save the searches. Your identity remains protected and your searches anonymous. This search engine stores neither an index nor a cache and instead simply reuses the index or results of one or more other search engines to provide an aggregated, final set of results.
- Integration with websites: Get fast and relevant search results. Customize the look of the search results to match your site's design.

VI. Future Scope

Broadly it can be effectively deployed in following real world instances:

- Helpful for one and all users who has to search any data in particular.
- Specific requirement is only been displayed which saves user's time.
- Various searches are been made which makes it optimized.

VII. Conclusion

Search engines have become an integral part. It is used in day to day lives of people. It is a search service that makes easier to dig out information from tons of resources available. It acts as a savior for people who are looking to save time i.e, information is available in a short span of time. Today, any person can search for number of differently phrased words or keywords related to specific topic and still come up with a plethora of information from great number of sources. Impact of search engines on popularity evolution of web pages is huge. Many search engines are being developed; organizations are coming up with some of the best improved search engines. Hence one can understand the importance of search engine technology. Mini search engine aims to find results according to the user's query, results that are usually quite accurate, presenting you with valuable information nuggets amidst a vast information mine. These search engines find specific information in few seconds making it more users friendly. Using 'OR' or 'AND' keyword will lead to more specific documents.

Hence the "Mini search engine "aims to serve the following points:

- To make searching of information more specific and precise according to user's requirement.
- To be time saving and make searching enjoyable for all.
- To handle wide range of specific types of information.
- AND and OR keyword makes searching more simple

REFERENCES

[1] "Seo for 2011: Search Engine Optimization Secrets
[2] "Word press 3 Search Engine Optimization: 2011
[3] An Introduction to Search Engines and Web Navigation, Mark Levene
[4] Search Engines: Information Retrieval in Practice, Bruce Croft, *,Donald MetzlerTrevor Strohman*
[5] Web Search Engine Research, Dirk Lewandowski
[6] GOOGLE'S PAGERANK AND BEYOND: THE SCIENCE OF SEARCH ENGINE RANKINGS BY AMY N. LANGVILLE, CARL D. MEYER
[7] Http://Searchengineproject.Wordpress.Com/
[8] Http://Codersview.Blogspot.In/2009/05/Mini-Project-On-Search-Engin.Html
[9] Http://Ieeexplore.Ieee.Org/Xpl/Login.Jsp?Tp=&Arnumber=781636&Url=Http%3a%2f%2fieeexplore.Ieee.Org%2fxpls%2fabs_All.Jsp%3farnumber%3d781636
[10] Http://Www.Slideshare.Net/Alisha1991/Search-Engine-9137278
[11] Http://En.Wikipedia.Org/Wiki/Pagerank
[12] Http://En.Wikipedia.Org/Wiki/Google_Search

