e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

IMPLEMENTATION PAPER OF SEARCH ENGINE USING WEB ANNOTATION

Mrs. Sonal Bawankule*[1], Ms. Aditi Bhaje*[2], Ms. Radhika Sangole*[3],
Ms. Arshiya Sheikh*[4], Ms. Prachita Darode*[5]

*[1]Professor, Department Of Computer Science And Engineering, Priyadarshini J.L College Of Engineering, Nagpur, Maharashtra, India.

*[2,3,4,5]Student, Department Of Computer Science And Engineering, Priyadarshini J.L College Of Engineering, Nagpur, Maharashtra, India.

ABSTRACT

The way of the people to access and find knowledge has effectively changed due to the search engines. It allows information about almost any subject to be quickly and easily improved within seconds. The content on the net is increasing day by day hence the influence of search engines on our lives will continue to grow. Search engine content analysis is a great evolution of the conventional data recovery field called collection selection. It deals with general information sources. To achieve high efficiency in web searching, the statistical method is proved to be an effective way. To design a search engine is a difficult assignment. The search engines like Google index tens to billions of net pages as concern to a corresponding extensive selection of awesome terms. They answer billions of queries every day. In spite of the importance of big scale search engines on the web, hardly few educational research has been done over them. Moreover, designing a web search engine at the present time is extremely different from three years before due to quick progress in technology and web creation. This paper will provides an thorough illustration of our big extent web search engine.

Aside from the difficulties of scaling standard search approaches to data of this magnitude, there are significant technical hurdles associated with exploiting the additional information inherent in hypertext to provide better search results. We concentrate on how to create a workable large-scale system that can take advantage of the extra information found in hypertext. We take a look at how to deal with unconstrained hypertext collections successfully. Anyone can publish anything they want on it.

Keywords: Web Pages, Keywords, Search Query, Search Engine, Web, Data.

I. INTRODUCTION

The web creates new challenges for information retrieval. The amount of information and new users unpracticed in the artwork of web research, is growing increasingly In general, individuals are accustomed to accessing the web using its connection graph. It's common to start with high-quality human-maintained recommendations like Bing or other search engines, the underlying software for automatically performing searches against a vast number of sources is known as search engines . A few advertisers are attempting to garner public attention by employing successful tactics to mislead automated search engines. We have created a wide-ranging search engine that will mark as many of the problems as possible of current systems. It makes maximum utilization of the extra formation which will available in hypertext to provide much high standard search results. We chose our system name personalized context search which fits well with our goal of building a search engine that gives personalized search results that are annoted.

1.1 Web Search Engines

Search engine technology has measured fiercely retrain with the widening of the web. The World Wide Web Worm (WWWWW), one of the first web search engines, contained a database of 110,000 web pages and web-reachable documents in 1994. The top search engines announced in November 1997 that they measure from 2 million web crawlers to 100 million online content. A full-scale index of the Web is estimated to contain billions of documents by the year 2000. Simultaneously, the vast majority of search engine requests were handled admirably. By the year 2000, major search engines will likely be handling hundreds of millions of inquiries per day, thanks to the growing number of web users and automated systems that query search engines. Our system's purpose is to address many of the issues, both in terms of quality and quantity.

1.2. Google: Scaling with the Web

There are numerous obstacles in developing a search engine that can grow to today's web. Fast crawling technology is required to collect web documents and keep them up to date. Storage space must be efficiently employed to store indexes and, presumably, the documents themselves. Hundreds of gigabytes of data must be processed efficiently by the indexing system. Hundreds to thousands of requests per second must be processed for each query.

1.3 Design Goals

1.3.1 Improved Search Quality

The primary goal of our project is to increase the quality of web search engines. Some people assumed that having a comprehensive search index would allow them to quickly locate anything. The greatest navigation service should make finding practically anything on the Internet simple and straightforward. People are still just interested in the first few outcomes. We need tools that can return a large number of relevant papers with a high degree of precision as the collection expands.

1.3.2 Academic Search Engine Research

Aside from phenomenal development, the Web has become increasingly commercial over time. In 1993,.com domains were used by 1.5 percent of web servers. In 1997, this figure had risen to almost 60%. At the same time, search engines have made the transition from academic to commercial use. Until now, the majority of search engine development has taken place within firms, with limited disclosure of technical data. As a result, search engine technology is still mainly black art and heavily influenced by advertising. We want to drive more development and understanding into the academic realm with Google.

Another significant design goal was to create systems that could be used by a large number of individuals. We were curious about usage since we feel that some of the most exciting studies will be based on a vast amount of data available from modern online technologies.

Our ultimate design goal was to establish an infrastructure that would facilitate cutting-edge research on massive amounts of web data. Before saving any documents, Google compresses them. One of our main goals when we started Google was to establish an environment where other academics could come in quickly and process huge chunks of the web., and come up with intriguing conclusions that would have been impossible to come up with otherwise.

II. METHODOLOGY

Search Engines use an algorithm and some set of rules to determine exactly which pages to show for any given query. These algorithms are complex and take into account 100 or 1000 different ranking factors for determining the ranking of search engines using web annotation. A search engine normally has the below-mentioned parts:

Web Crawling

Web crawling, also called ‘Web Spider’. It's a program or script that goes over the World Wide Web (WWW) in a systematic, automated manner. It can be used by a search engine to provide up-to-date information and create a copy of all the visited pages to be stored in an indexer that helps to index these pages to speed up the search process. The automated script, on the other hand, can be used to automate maintenance activities such as checking links and validating HTML codes, or even gathering specific information from websites, such as email addresses, and phone numbers within special areas. It will begin with its seeds, which contain a list of URLs, and then identify all of the hyperlinks on the page, which it will then add to the seeds, which will be visited recursively according to some selection principles.

Web Indexing

To support the information retrieval of efficiency and accuracy, the search engine indexer collects, parses, and stores data to optimize performance for the search query. The majority of prominent search engines concentrate on full-text indexing of internet documents written in natural language. After a web crawl robot obtains information, the search engine indexes the data according to the algorithm and stores the indexing file in a database. The next generation of search engines aims to develop more intelligent parsing and indexing technologies that could understand and satisfy people’s requests in natural language.

Searching Mechanism

The searching software part searches information from the indexer based on a search query and returns the matched results. Using web annotation, the matched results are then annotated. The search results appear in a higher index format using the page rank algorithm.

Page Ranking

[1] The success of Google is attributed to the way it ranks web pages, i.e. the ranking algorithm to measure the importance of a web page. Quoting from the original Google paper, Page Rank is defined as (Brin & Page, 1998): "We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor that can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also, C(A) is defined as the number of links going out of page A. The PageRank of page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Because Page Ranks form a probability distribution over online pages, the total Page Ranks for all web pages will be one. PageRank, also known as PR (A), is a simple iterative method that calculates the primary eigenvector of the web's normalised link matrix." It is fully determined by the structure of the World Wide Web.

III. MODELING AND ANALYSIS

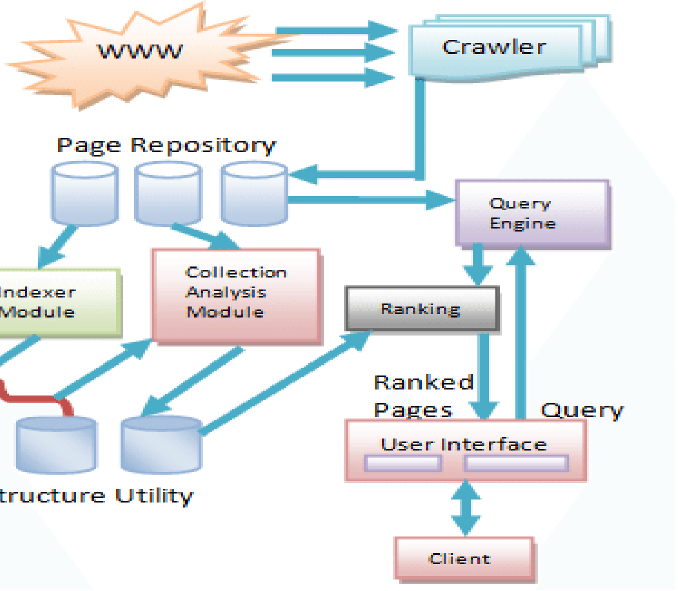
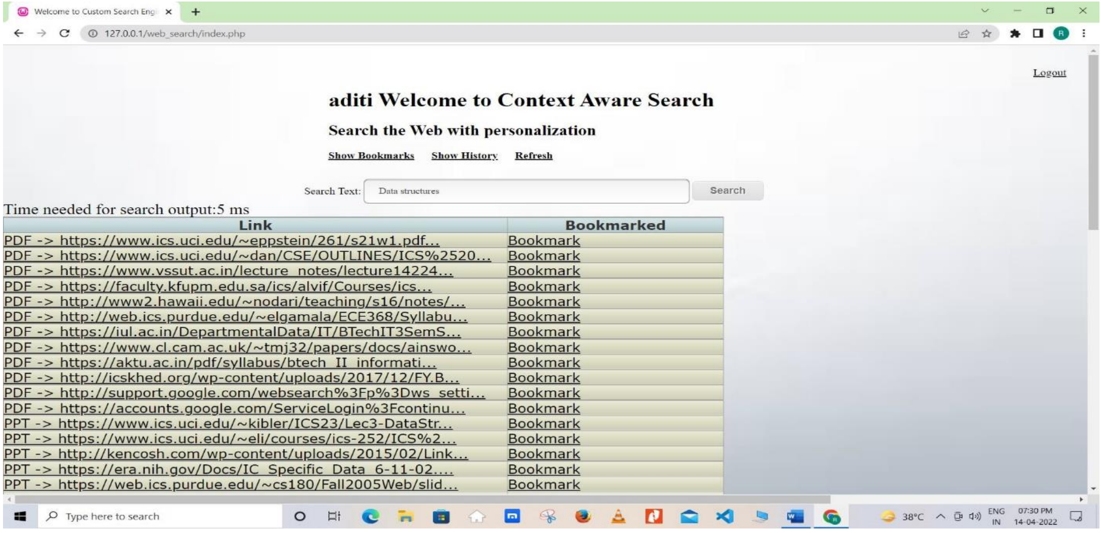


Fig: Basic Search Engine Architecture

IV. RESULTS AND DISCUSSION



aditi Welcome to Context Aware Search

Search the Web with personalization

Show Bookmarks Show History Refresh

Search Text: Data structures Search

Time needed for search output:5 ms

Link	Bookmarked
PDF -> https://www.ics.uci.edu/~eppstein/261/s21w1.pdf...	Bookmark
PDF -> https://www.ics.uci.edu/~dan/CSE/OUTLINES/ICS%2520...	Bookmark
PDF -> https://www.vssut.ac.in/lecture_notes/lecture14224...	Bookmark
PDF -> https://faculty.kfupm.edu.sa/ics/alvif/Courses/ics...	Bookmark
PDF -> http://www2.hawaii.edu/~nodari/teaching/s16/notes/...	Bookmark
PDF -> http://web.ics.purdue.edu/~elgamala/ECE368/Syllabu...	Bookmark
PDF -> https://iul.ac.in/DepartmentalData/IT/BTechIT3SemS...	Bookmark
PDF -> https://www.cl.cam.ac.uk/~tmj32/papers/docs/ainswo...	Bookmark
PDF -> https://aktu.ac.in/pdf/syllabus/btech_II_informati...	Bookmark
PDF -> http://icskhed.org/wp-content/uploads/2017/12/FY.B...	Bookmark
PDF -> http://support.google.com/websearch%3Fp%3Dws_setti...	Bookmark
PDF -> https://accounts.google.com/ServiceLogin%3Fcontinu...	Bookmark
PPT -> https://www.ics.uci.edu/~kibler/ICS23/Lec3-DataStr...	Bookmark
PPT -> https://www.ics.uci.edu/~eli/courses/ics-252/ICS%2...	Bookmark
PPT -> http://kencosh.com/wp-content/uploads/2015/02/Link...	Bookmark
PPT -> https://era.nih.gov/Docs/IC_Specific_Data_6-11-02....	Bookmark
PPT -> https://web.ics.purdue.edu/~cs180/Fall2005Web/slid...	Bookmark

Fig: Result for the Query

The necessary notion of a search engine is the attribute of search outcomes. As in the below figure, we can see that the query searched is the data structure and we obtained the links of the results in a sorted format. The available PDFs are appearing first then all the PPTs format documents are appearing, it shows the sorted result after searching the query using the Query matching algorithm The maximum results are high-quality pages. The pages are of excellent quality as they all have high PageRank. The user can mark the important links as bookmarks and it'll show up in the bookmarked links section and the user will be able to access them easily. The users can also check the queries which have been searched by them by simply clicking on the Show History button. It will provide efficiency in the search results and the user will obtain all the data in the sorted format.

V. CONCLUSION

In this research paper, a discussion on how to add value to search results collected as a result of querying or searching the Web was made. The search engine using web annotation focuses on the matter of web searching and web search engines. It comprises several parts, that will always follow the rule of practice which is supported by theoretical background. In this paper, the author introduces the basics and most important topics related to the internet. And web searching also. We have suggests a substructure for a personalized web search which will be used to creating an Enhanced User Profile using browsing history.

In this paper, further research was done on how to add value to the results brought about by querying the internet. Two aspects were viewed: Annotating search results to add value to them and categorizing search result records for simplifying identifying relevant search result records. An implementation of a search interface was made that consisted of a Web Crawler for fetching and indexing Web documents, a Search Engine for searching and displaying results, and a Classifier for categorizing search result records. The results showed an improvement on earlier work, where a tool for collecting data in the competitive intelligence field can have search results well-annotated and classified.

VI. REFERENCES

[1] https://www.ccs.neu.edu/home/vip/teach/IRcourse/4_webgraph/notes/Pagerank%20Explained%20Correctly%20with%20Examples.html

[2] Yahoo! http://www.yahoo.com/

[3] http://www.searchenginewatch.com/

[4] ftp://ftp.uu.net/graphics/png/documents/zlib/zdoc-index.html

[5] http://harvest.transarc.com/

[6] http://google.stanford.edu/

[7] http://info.webcrawler.com/mak/projects/robots/exclusion.html

[8] Witten 94] Ian H Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.

[9] [Weiss 96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.

[10] Witten 94] Ian H Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.

[11] [Weiss 96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. pursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.

[12] http://botw.org/1994/awards/navigators.html