Implementation of a Search Engine

Ashish Kumar Garg[1], Mohammad Amir[2], Jarrar Ahmed[3], Man Singh[4], Sham Bansal[5]

[1]Delhi University, Department of Mathematics, JDM College, Rajinder Nagar, New Delhi-110060, India

[2]Delhi University, Department of Mathematics, IP College for Women, Civil Lines, New Delhi-110054, India

[3]Delhi University, Department of Mathematics, Dyal Singh College, Lodhi Road, New Delhi-110003, India

[4]Delhi University, Department of Mathematics, SPM College, West Panjabi Bagh, New Delhi-110026, India

[5]Delhi University, Department of Mathematics, Bharti College, C4, Janakpuri, New Delhi-110058, India

Abstract: ***Today's world the information is most valuable quantity. With the advent of the web, the information storage and retrieval have taken a huge step forward. Search engines plays important role in this area. In this report (implementation of a search engine), we talk about the functionality of a mini-offline search engine. We study the various components of search engine is which involves "crawlers" (a spider program to search through documents), "porter and stemmer" (program that remove stop words and brings the query in its basic form) and "indexer" (one which indexes the documents to cut short the duration of searching). Next part is the implementation of these various components .The search engine while searching through the web gives us so many relevant and irrelevant. Which relevant information should come as a best desired result, depends on the kind of algorithms that all of the search engine have got the propriety right over. I have also tried to develop a similar algorithms named as page rank, which has been implemented on a small web graph and extract the information through the in-links and out-links of any web page.***

Keywords: Search Engine, Page Rank, Web Mining, Links, World Wide Web

1. Introduction

Information retrieval is finding material of text, images, audio and video that Satisfies information need from within large collection. In today's world with the advent of the web, information retrieval and storage have grown many folds. Here we would like to extend this information retrieval to the web search. When we talk about the web search, the system has to provide search over billions of documents Stored on millions of computer systems [1]. The complexity of designing these systems Is being able to build systems that work at enormous scale like being able to store and search through documents. The systems not only are user friendly but also should be responsive in the very short span of time. An information retrieval begins when a user type any query into the system. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several collections may match the query with some degree of relevancy (Search Ranking.).

2. Use of Search Engine in Information Retrieval

The World Wide Web has become a global information space of enormous size. Information Retrieval is about finding material of text, images, and audio and video that satisfies an information need from within large collection. In today's world with the advent of the web, information retrieval and storage have grown many folds. Here we would like to extend this information retrieval to the web search. When we talk about the web search, the system has to provide search over billions of documents stored on millions of computer systems [2].

The complexity of designing these systems is being able to build systems that work at enormous scale like being able to store and search through documents [4]. The systems not only be user friendly but also should be responsive in the very short span of time. An information retrieval begins when a user type any query into the system. In information retrieval a query does not uniquely identify a single its mass scale, heterogeneity, distribution and dynamic characteristic cause information overload. How to retrieve web information effectively in order to location and obtain the information we need with great speed become an important and urgent issue. Search engine is a kind of information retrieval tools adapting to web characteristics on the basis of traditional information retrieval techniques. It finds and collects information on web on its policy, after understanding, processing and organizing this information, provides web information retrieval services for users, playing the role of information navigation. According to information collection and service delivery mode, we mainly classify these search engine systems two types as follows:

2.1 Robot Based Search Engine

These search engines traversal web in a certain strategy using software robot, which is also known as a spider or a crawler, download web documents to the local document libraries for analysis, build up index by the indexer, retrieve the index database in accordance with query requests accepted from user interface, and finally output the query results to users. They own a huge full-text indexing database, mass information and high frequency of updates, are suitable for retrieving artificial information request. But their results are too many their information accuracy is very low, we must filter these results to find information we need [6].

2.2 Directories Based Search Engine

They collect web information by artificial collection or website authors' initiative commitment, review, categorize,

and summary these websites and documents and place them into the classification framework, and organize resources in tree structured directory classified by subject [18]. We can not only visit this directory downward stage by stage from the root classification until finding what we need, but also submit our request direct and let the system help us to location the aim resources. Introducing human intelligent, their retrieve results are more accurate and navigation quality better. But manual mode cramps its frequency of updates and content capacity [9]. As there are respective advantages and disadvantages in both robot-based and directory-based search engines, many search engine systems provide robot-based and directory-based retrieve service at the same time to make their results more accurate and comprehensive.

3. Web Mining

Before we study the search engines it is imperative to have an understanding of Web Mining. Web mining is application of data mining techniques to extract knowledge from Web data. The Web data comprises of Web content, Web structure, and Web usage [10]. We can use web mining to do resource finding, information selection and pre-processing, generalization and analysis. Web mining is the application of data mining techniques to extract knowledge from Web data. Before going into the details of web mining we take a look at the WEB.

3.1 The Web

The present web graph has an estimated 150 million node and 1.7 billion edges. The World Wide Web is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them via hyperlinks. Using concepts from earlier hypertext systems, British engineer and computer scientist Sir Tim Berners-Lee, now Director of the World Wide Web Consortium, wrote a proposal in March 1989 for what would eventually become the World Wide Web. "TheWorld-WideWeb was developed to be a pool of human knowledge, and human culture, which would allow collaborators in remote sites to share their ideas and all aspects of a common project. The web graph describes the directed links between pages of the World Wide Web [5].

A graph, in general, consists of several vertices, some pairs connected by edges. In a directed graph, edges are directed lines or arcs. The web graph is a directed graph, whose vertices correspond to the pages of the WWW, and directed edge connects page X to page Y if there exists a hyperlink on page X, referring to page Y. Web mining makes use of this graph. Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

3.2 Web Usage Mining

Web usage mining is a process of extracting useful information from the user interaction with the web. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data such as text files; pdf files etc, whereas some others might be interested in multimedia data which include pictures, videos, music etc. Web usage is discovering new, interesting, and meaningful patterns generated by the users during the client-server transactions on one or more web localities [8].

3.3 Web Content Mining

Web content mining is the process of discovering useful information from text, image, audio or video data in the web, i.e., the information conveyed from the actual content of the web page. Web content mining sometimes is called web text mining, because the text content is the most widely researched area. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval). Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

3.4 Web Structure Mining

Web Structure Mining is the extraction of information from the underlying structure of the web, i.e., the web graph. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage. The hyperlink structure of the web helps in: Navigational purposes. Relevance of pages [12]. This can be used to retrieve useful information from the web. It is of importance that we study some terminology related with the web structure. Node: each page in the web graph is known as a node Link: the hyperlinks form the directed edges between the nodes In-degree: the total incoming links or pages pointing to a page Out-degree: the total outgoing links that the page points to Directed path: a path staring on page `a' that can follow links to page `b 'Now that we are familiar with the web graph, let us have a look at some Interesting web Structure [16].

4. Page Rank

Search Engines are the most important providers of information on the web and Google has been leading player in the area in the recent few years. The reasons can primarily be attributed to the fact that its results are accurate and comprehensive [17] . The Google Page Rank algorithm is one of the most important reasons for the same. The internet can be viewed as a large graph with pages corresponding to nodes and links as edges. The Page Rank algorithm decides how important a page is and hence where it will show up in the search results. The main idea behind the algorithm is simple: a page is important if it has a large number of other pages pointing to it, i.e., it has a large number of back links from other pages.

5. Algorithms

Step1. PR0 taken at arbitrary
Step2. Loop
PRi+1=P*PRi

Compute the effect of Random Walk and Go to Step2 until some convergence.

The above algorithm gives a very primitive solution for computing Page Rank through the power method.

6. Conclusions

This report introduced the functionality of a Search Engine. our implementation of the parts of Search Engine have been developed and coded. We have used the Linux as part of implementation. The page rank introduces the concepts of giving importance to the web pages that a search engine has to provide. Performance of this page rank algorithm was tested on web pages and has the capacity to be improved further with sparse matrix concepts.

7. Future work

A major application of page rank is searching. As I have done study on how this page rank can be implemented with search engine in order to get the better result of the searched queries. The benefits of page rank are the greatest for underspecified queries. For example, a query for IIT Kharagpur may return any number of web pages which mention IIT or Kharagpur (such as publication lists) on a conventional (simple title-based search engine)search engine ,but using page rank, the IIT Kharagpur home page is listed first .then assembling various components of Search Engine with Page Rank is also to be done. Further, with the help of CGI programming we can make it to function as a mini search engine.

References

[1] Christopher D.Manning, Prabhakar Raghavan, Hinrich Schutze, Introduction of Information Retrieval, 2008.
[2] Tang Jiutato,Lin Guoyuan, Shaung Xiaochaun, Research and Development of Search Engine Technology, Energy Procedia,(2011).
[3] Gulli A, Signorini A,The Indexable web is more that 11.5 billion pages,14th International World Wild Web Conf.(2005).
[4] Jaideep Srivastava, Web mining: accomplishments and future direction.
[5] Berners Lee, Tim Mark Fischetti, Weaving the Web: the original Design and Ultimate Destiny of the World Wide Web by its inventor, 1999.
[6] Dai Jing, Wang Yan Zhou Xuan, Web structure mining.
[7] Mei Kobayashi and Koichi Takeda, Information Retrieval on the Web, 2008.
[8] Larry Page, Sergey Brim, Motwani, The page rank citation ranking: bringing order to the web, 1998.
[9] Sergey Brin and Larry Page. Google search engine http://google.stanford.edu. , 1998.
[10] Jiawei Han and Micheline Kamber , Morgan Kaufmann , Data Mining: Concepts and Techniques, 2001.
[11] Bing Liu, Web Data Mining Exploring Hyperlinks, Contents , and Usage Data, First Edition , Dec 2006, Springer,2011.
[12] Marios D. Dikaiaosa, Athena Stassopouloub, Loizos Papageorgious, An investigation of web crawler behavior: characterization and metrics, computer communication (2005) 880-897.
[13] James E. Pitkow, Charcterizating World Wide Web Ecologies PhD Thesis, Georgia Institute of Technology 1997.
[14] Peter Piroli, James Pitkow, and Ramana Rao, Extracting usable structure from web, Proceeding of the Conference on Human Factor in Computing System 1996.
[15] Ron Weiss, Bienvenido Vielez, Mark A. Sheldon, Chanathip Manprempre Peter Szilagyi, Andrez Duda ,and david K.Gifford ,HyPursuit: A hierachicak network search engine that exploits content-link hypertext clustering .In Proceeding of the 7th ACM Confrence on Hypertext 1996.
[16] Ellen Spertus, Parasite: Mining structural information on the web, In Proceeding of the sixth International WWW conference, santa claram USA 1997.
[17] Sougata Mukherjea, James D. Foley, and Scott Hudson, Visualizing complex hyper media networks through multiple hierarchical views, in proceeding of ACM CHI'95 Confrence on Human Factor in Computing System 1995.
[18] Jon lienberg, Authoritative source in a hyperlinked environment, In Proceeding of the Nineth Annual ACM-SIAM Symposium on Discete Algorithms 1998.

Author Profile

Ashish Kumar Garg received the M. Sc degree in Mathematics from Indian Institute of Technology Madras and M. Tech. degree in Computer Science and Data Processing from Indian Institute of Technology Kharagpur 2013, India.