Survey Paper based on Search Engine

[1]Sathya M., [2]Reka S.

[1,2]Assistant Professor
Mount Zion College of Engineering and Technology

***ABSTRACT*: The purpose of this survey is to study the working of search engine using search engine optimization, web crawler and web mining. The large growth of World Wide Web (web) needs an efficient system that will manage the data over a web. This huge data get generated, deleted, manipulated every day and it is a source for millions of people all over the world. Search engine is the system provides relevant presentation of data that is to be searched. Search engine show the working of search engine optimization, web crawler, web mining. Search engine works on the architecture of web crawler that uses the web mining as well as search engine optimization to gather and represent the related data.**

***Keywords*: Web crawler; web mining; search engine optimization; search engine; World Wide Web (WWW); URL normalization; URL (Uniform Resource Locator)**

1. INTRODUCTION

World Wide Web (Web) is a dynamic system that rely on its information system which is get interlinked through hyperlinks or URL(Uniform Resource Locator) that is access via internet usually called as web. Web is a client-server system. In this system, client is a user that used to interact with the remote machine called server. Client used to request remote server for a data that is to be download via internet thus in response server provide the data in the application level of client. Client uses application for the interaction with remote server known as web browser i.e. provides facility to use the internet.

The massive growth of web in which huge amount of data collection get reside in the form of hypertext, text, media content etc. this data are highly unstructured thus to access this data efficiently, the framework is required that manages the data as well as presentation of data as per the requirement of user. Mostly data are represented in the form web pages. These web pages are the web document which is written in Hypertext Markup Language (HTML) that is used to present the data.

Search engine is a system that is introduced to search the data content of web. Search engine architecture relies on the architecture of web crawler. Search engine is a popular system so it has to be very efficient in terms of response time and database thus it make use of web mining and search engine optimization to present the searched data e.g. user use keyword 'computer' for searching thus it using the keyword to gather the related search using data mining then all data get arranged in list with their URL in a way that is part of search engine optimization. Every time click on URL, hypertext result in downloading of new web page that is part of web crawler.

SEARCH ENGINE OPTIMIZATION

OVERVIEW

Selection of valuable data from the set of huge available data this is referred as optimization. Search engine optimization is a way to show the extracted data on the basis of some set of rules that is followed by search engines. Search engine follows some instruction like paid website should be shown at top of unpaid website, website with high popular rating and maximum number of visitor has to be shown at top and so on. Also it provide an interface w.r.t the category of content like image, video, place wise search, documents etc.

Workings of search engine rely on SEO that provide terms and conditions for keywords that will be going to search. SEO has to keep track of ranking of website, estimate traffic, main alternative text of web pages, main referrers etc. The results are generally presented in a form of list of results also referred as SERPS (search engine results pages). Basic task of SEO is to manage the task performed by web crawler and web mining.

WORKING

First, search engine crawl to get the content fetched and this is performed by web crawler. Crawler used to traverse the URL's to retrieve the data (text, Meta data etc.) and save it to database also it helps to get the list of URL's that will be going to search. Second, after web page gets crawled, the next step is to index retrieved content and this index or log gets stored in large database. Indexing is done to find the related searched data easily by comparing the keywords with the Meta data stored. Indexing help to save the time to traverse large database and to find accurate data.

Fourth, when user searched for any keyword that get compared with the indexed pages in the database. Search engine compare it and provide the related links which is suitable for user aspects. This is done by the algorithm that followed by SEO there are some factors on which algorithm rely: keyword density, hyperlinks, Meta tags and so on. Keyword density refers to the factor that presents the number of that keyword get searched or a popular searched keyword. Last step is to retrieve the results and delivered it to web browser or end user.

2. WEB CRAWLER

OVERVIEW

Web crawler is a system that used to retrieve the URL's information. It helps to traverse from web pages by seed the URL. It traverses around the internet and create an index of all collected information. When all data get retrieved from URL's then it uses the external URL’s within it to attend the other web page. Web crawler is major component of search engine and other application that process large numbers of web pages e.g. e- commerce sites and web data mining and so on. Web crawler have to process thousands of pages per second behind this simple description a bunch of issues get related to the network connection, spider trap , canonicalizing URL’s, parsing HTML pages, ethics of dealing with remote server and so on.

High efficient web crawler is required to handle the millions of web pages index in search engine. In fact there is a high compete between various search engines in terms of size and currency of their underlying database, quality of response time. Thus web crawler has to work on some issues. First, web crawlers have an efficient algorithm, i.e. web crawler should be able to decide which web page is to be retrieve next. Second, web crawler report should be updated in a short span of time. Third, two different web crawler have to share a cache thus to avoid requesting of same web page by both web crawler. Fourth, web crawler has to be highly efficient system thus to retrieve large number of pages per second while being robust against crashes, web server related issues and so on.

3.1 WORKING

Web crawler response time basically not only depend on internet connection but also by data of the web page that is to be crawled. If a crawling is done for a web page from multiple server i.e. if many downloads are performed parallel. Then that will reduce response time of crawling. Thus web crawler is implemented in multi-threading to perform task parallel.

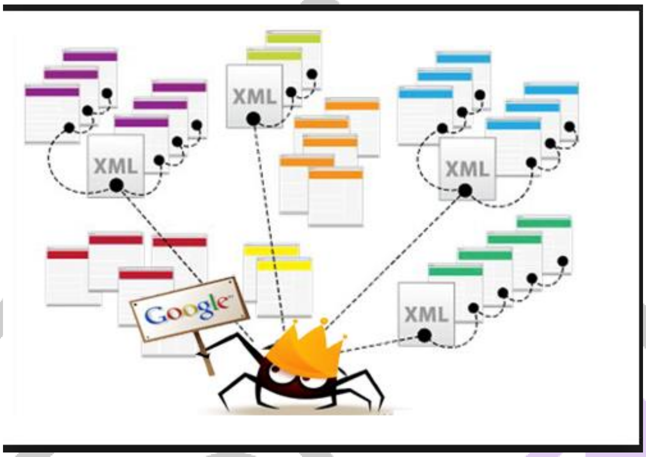


Fig (1) Web Crawler Architecture

Firstly, web crawler start with the single URL called as seeds before attending the URL it perform URL normalization. URL normalization is a process in which URL get transformed into normalize or canonical URL thus to overcome different URL’s that seems have similar text e.g. the URLs http://google.com/news and http://news.google.com return similar data from the remote server.

Second, it fetch all hyperlinks within the web page and add it to the queue called as frontier URL that act as a to-do list for the crawler that has been unvisited. Crawler retrieve the information of web page and save text related data to database i.e. website text content, Meta tag content and so on. It doesn’t retrieve any media content i.e. flash, animations and any dynamic resources.

Third, while attending the hyperlink this loop is performed until all hyperlinks from frontier get visited once. Scheduler help to schedule the frontier URL i.e. to be next visited

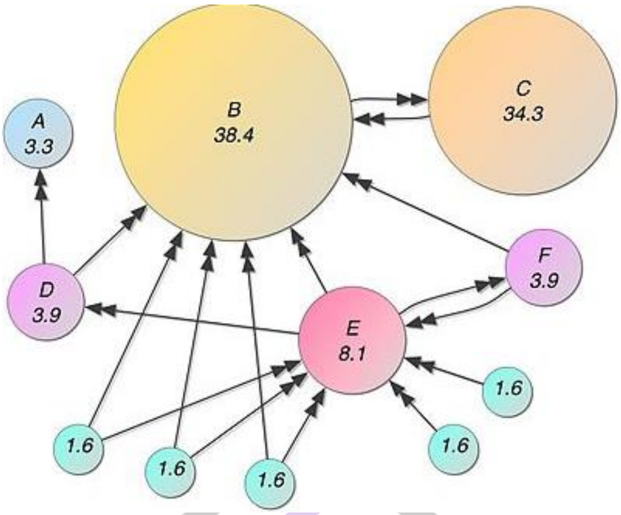
3.2 SEARCH ENGINE OPTIMIZATION

It is the way of increasing the visibility of a page by natural means i.e., unpaid search results. In this process the website undergoes redevelopment to make our keywords effectively communicate with major search engines. This work is done by SEO (Search engine optimizers), They may target image search, academic search, local search, video search. Optimising a page involves editing contents & HTML codes in order to increase its relevance to specific keywords and proper indexing in search engines .The contents and codings are edited keeping in view of the indexing pattern of the search engines which are done by a crawler named Googlebot in Google. It is the most powerful way to reach to reach the customer as we meet them when they are in need. Most of the users find the target websites during their search.

PageRank

It is an algorithm used by Google which assigns numerical weight to the URL of web documents to measure its relevance. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by PR (E) [1]. Stanford University is the birthplace of PageRank when Larry Page (hence the name Page-Rank) and Sergey Brin were involved in research of a new kind of search engine. The idea of Sergey Brin was that information on the web could be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it. In 1998, the first paper describing the PageRank and initial prototype was published after which Page and Brin founded Google Inc., the company which is behind the Google search engine.

It shows the popularity or a particular link or a website. The page with higher rank gives more optimised results. PR(A)=(1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) [2] The name "PageRank" is a trademark of Google, and the process has been patented (U.S. Patent 6,285,999). The said patent is of Stanford University to which Google has exclusive license rights. The university received 1.8 million shares of Google in exchange for use of the patent; the shares were sold in 2005 for $336 million [3].



REFERENCES

[1] Configuration File of W3C http, 1995. http://www.w3.org/Daemon/User/Config/.
[2] S.Balan and Dr.P.Ponmuthuramalingam, “A Study on Semantic Web Mining And Web Crawler”, IJECS, vol. 2, pp. 2659-2662, Sept 2013.
[3] Subhendu kumar pani, Deepak Mohapatra and Bikram Keshari Ratha, “Integration of Web mining and web crawler: Relevance and State of Art”, vol. 2., pp.772-776, 2010.
[4] Rajesh Singh and S.K.Gupta, “Search Engine Optimization
- Using Data Mining Approach”, IJAIEM, vol. 02, ISSN 2319-4847, sept 2013.
[5] Rajesh Singh and S.K.Gupta, “An approach for Search Engine Optimization Using Classification - A data Mining Technique”, IIJCS, vol. 02, ISSN 2321-5992, April 2014.
[6] Gautam Pant, Padmini Srinivasan and Filippo Menczer, “Crawling the Web” unpublished.
[7] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[8] Kulvinder Singh, Yogesh Kumar, “Search Engine Optimization using Sequential Pattern Mining and Page Ranking,” IJSWS, ISSN(Online) 2279-0071.
[9] Ricardo Baeza-Y ates, Web Mining in Search Engines, Center for Web Research Department of Computer Science Universidad de Chile Blanco Encalada 2120, Santiago, Chile
[10] Marc Najork, “Web Crawler Architecture”, Microsoft Research, Mountain View, CA, USA
[11] Monika yadav and Pradeep Mittal , “International Journal of Advanced Research in Computer Science and Software Engineering”, IJARCSSE,vol. 3, issue 3, ISSN: 2277 128X, March 2013.
[12] Joeran beel, Bela gipp and Erik wilde, “Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.”, University of Toronto Press, January 2010.