Abstract

The processing of extracting information is too much costly in terms of processor time. A distributed design approach is proposed in this paper. Advantage, limitations are also discussed for this approach. Our main objective in this paper is to design a such type of search engine architecture in which processing of a search engine may be in distributed form by which the time or processing power may be reduced.

Keywords: Web Crawler, Structure, Policies, Cluster.

Introduction:

World Wide Web provides us with huge amount of necessary data digitally available as hypertext. Data may be WebPages, images, information and other type. This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information. So Web Crawler for automatic Data and Web Mining is Useful to Us.

The economic importance of web will enhance the academic interest. The Database Administrator, Management persons or others wishing to perform data mining on large number of web pages will require the services of web crawler or its based tools. For these reasons crawlers are normally multi threaded by which millions of WebPages may be extracted parallel by only one process.

This paper shows distributed approach for web Crawlers including data mining. It works as a distributed system with a central Control Unit fixing or providing the jobs to different computer. Which are connected with a Network? This proposal is not exactly new as large search engine can also distribute processing power by using different number of computers Systems linked with each other.

Search Engine

Search Engine provides the gateway for most of the users trying to explore the huge information base of web pages. Search engines are programs that search documents for specified keywords on search for information on the World Wide Web and returns a list of the documents where the keywords were found. A Search Engine is really a class of programs; however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web.

Goals of Search Engine:

(1) Quality: Means effectiveness can be defined as to retrieve the most relevant set of document for a query. Process text and store text statistics to improve relevance be used.
(2) Speed: Means efficiency may be defined as a process queries from users as fast as possible For it specialized data structure should be used.

How web Based Search Engine Works?

Web based search engine works by saving the information of many web pages, which they retrieve itself. These pages are retrieved by a web crawler which is also called spider which follows every link on the site (Figure 1[3])

Search engine is a term used for information retrieval. Search engine match queries against an index that they create. This index contains the word in each document, pointers to their location within the document. This is called inverted file.
Features of Web based Search Engine:
Following are the basic features for evaluating web based search engine-

1. Web Indexes: When a web search request is generated, it is the web index generated by web robots or spiders. The combination of web indexes affects the performance of a web search engine. Three main key points to design of web index are coverage, update frequency and the part of indexed web page.


3. Retrieval Issue: This issue proceed on three Key points- Precision, Recall and response time

4. Write Option: Write option or output option provides the deal with actual content of output.

5. User effort: User effort means the documentation and interface. Good prepared documentation and good interface play a different role in users’ selection of web search engine. User will only use the search engine when the interface is user friendly only.

Quality of Good Search Engine:

1. Ability to produce the most relevant result to any given search.

2. A true search engine is an automated software program that moves around the web collecting WebPages to include in its catalog or database.

3. It searches when user requests information from a search engine has its own catalog or database of collected WebPages, so you will get different results. Hits by using different search engines.

Problems Facing by Current Search Engines:

1. Crawlers are not able to analyze the content of keyword in web page before they download it.

2. User submits his request for retrieval of information without mentioning the content in which he otherwise desire.

3. Crawler treats user search request in isolation.

4. There is a requirement to prepare separate files for each web document.

5. Augmentation is required in HTML document.

Types of Search Engine: According to functioning three types of search engine [7].

1. Crawler Based Search Engine:
They create their listings automatically. Spider builds them. Computer algorithm ranks all pages. These types of search engines are heavy and often retrieve a lot of information. For complex search it allows to search within the results of previous search and enable you to refine search results.

2. Human Power Directories:
These are designed by human selection means they depend on professional to create listings. These never contain full text or webpage they link to.

3. Hybrid Search Engine:
These are different from traditional text oriented search engine such as Google or directly based searched engine such as Yahoo in which each program operates by comparing a sets of metadata.

Search Engine Optimization:
Search Engine Optimization is the procedure of improving the visibility of a website or webpage in search engine via the natural or unpaid searched results. Optimization may target different types of search like image search, local search, video search, academic search, new search, industry specific vertical search. It can also be define as the process of affecting the visibility of a website or webpage in search engine.

In search engine optimization updations or modification of all variables to get a better location in the search engine take place. We start with Sarch Engine Optimization and how it can be used to formulate internet marketing. Strategy as well as Technical aspects of SEO. [4]

A) Using SEO as a marketing strategy it can be described as a method of getting our website to rank higher in search engine as Google; Yahoo,
Means that if user is liking to search for a list of optimized keywords the chances are that the visitors see your site on first few places may be good.

B) Parameters for evaluating SEO of websites- Page Rank- Page rank of each page depends on the page rank of pages pointing to it.

C) To enhance our site page rank few key ideas are inbound links, outbound links, Dangling links, domain and File names and Broken links [4]

Search Engine Optimization Technique
Basically three techniques for search engine optimization are there [6].

1. Directory Submission:
It is the important technique in Search Engine Optimization to create incoming links to a website through indexed page and category. Different directory provides free service to website. Directory submission request information regarding URL, title, keywords.

2) Keyword Generation:
All search engine optimization need some words to elaborate information based on these words. Keywords should be of your organization on subject. This process can be proceeding by different online tools like word tracker, yahoo keyword selector tool, Google Ad words.

3) Link Exchange:
To start up any website for any business we need reciprocal link exchange with other websites. It is the procedure to take place link on other website and other website place links on our site.

Tools of Search Engines Optimization:
SEO tools are the operators that optimize the search engine functionality.[5] Basics tools are-
1) Keyword Tool- Include keyword research tools, keyword density analysis tool, and competitor analysis tool. It can used for website classification and regulate keywords deployment columns. Eg – Keyword selector tool, keyword external tool
2) Link Tool- These tools include link popularity spider simulator, by which ranking of website can be increased.
3) Usability Tool- This tool test pages display effects in different resolution, different operating system, and different browser. These include HTML and CSS validation, Firefox extension, and Page speed test.
4) Keywords Strategy- When choosing keywords, it must be relate with products, area, service.
5) High duality incoming Link- Submit the website to search engine directories, find websites to exchange links. In it import link, outbound links, internal link are used.

Criteria for site optimization
For a new website to be optimized for the given keywords need to have some technical issues checked.[2]
1) Meta descriptions or Metadata Keywords
2) Keyword analysis
3) Title Tags
4) Page content
5) Headlines Tag
6) URL structure and domain
7) Images Tag
8) Page Load time
9) XML site Map
10) Meta data using schema
11) Site map
12) Robot.txt
13) 404 error
14) Duplicate contents

Crawling Techniques:
1) Focused Crawling:- Focused crawler is designed only to retrieve documents on a specific topic, thus reducing the amount of network traffic and downloads. The goal is to selectively seek out pages that are relevant to a pre defined set of topics. This leads to savings in Hardware and Network resources and helps keep the crawl more up to date.
2) Distributed Crawling:- A single crawling process is non-useful for large scale engine that needs to fetch large amount of data rapidly. Distributing the crawling activity via multiple processes can help build a scalable, system which is fault tolerant system. Distributing the load decreases hardware requirements and at the same time increases the overall download speed and reliability.

Proposed Architecture:-
This architecture should be capability of systems working in a Distributed manner. In it all processing should be processed on idle computer. The distributed architecture should not increase Network traffic. All the systems connected in a Network should be operated using some firewalls. The module should be as much as easy to plug and play.
1) Crawler Unit:- It will crawl a website. It will need to use secondary memory to store the web pages downloaded before analysis. The web pages should be saved on each host system, rather than transferred to the control unit to minimize the network traffic. In crawler unit the technique of Data mining that is cluster may be applied by which similar data elements, similar URL’s may be kept as a cluster. Cluster helps us to crawl the different pages. Different types of clustering algorithm may be used to crawl the useful URL’s.
(2) Control Unit: When a crawler requests a job or sends some data elements, the control unit will live on a web server and will be used by it. It will need to save the commands that the user likes to be processed. This can also be understood by this example as data files are saved on a server.

(3) Messaging System: To satisfy the necessity of crawler and control unit, the crawler must be able to download and process websites with less transmission with control unit. The crawler unit starting a crawl, sending a message to control unit shows that it is ready to execute a new request. The control unit sending an instruction to the crawler shows a site to crawl and type of processing to be performed on downloaded websites.

The problem is that control unit outside the network can not initiate communication with component inside but only can send information in response to a request for it as per Figure 2 [8]. This architecture can be successful for any component where the control unit is on a public access web server. The architecture described here is employed to design a system for the job of analyzing the link structure of websites. This program had not run quickly enough to consider necessary number of websites and so it has been individually setup and run on a number of computers parallel.

One more feature that was built into the crawler was an option of types of checking for duplicate pages to be used in a website crawl.

There are three options:
- Use no page checking think that two page with different pages are different pages.
- Use HTML page checking oppose new page which is identical HTML to later retrieved page.
- Use weak HTML Page Checking.

Clustering is a method in which like records are grouped together. This technique is done to give the end user a high level view of what is going on in the data set. Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest.

Clusters may be created either statistically or by using artificial intelligence methods. Clusters can be analyzed automatically by a program or by using visualization techniques as described in fig-3 [11].

Future Work:
Search engine is a complex system on which further enhancements should be made. Some of the key ideas are like using query caching, disk allocation, sub indices, RAID techniques. More advanced algorithms are also required to decide which old pages should be re crawled and which should be new one crawl. Normal features by commercial Search engine like Boolean operators, negations, streaming use of AI should be added.

Conclusion:
The architecture defined here is capable of crawling a large number of websites. It can not process 100% automatic for jobs that involve crawling entire websites without heuristic for finding or searching duplicate pages. This design approach is suitable for the situations where a job can be divided into a disconnected crawling based jobs by which execution on different systems should not produce a problem.

It may be non-useful if the crawls have to cross transmission each other in any case, for example to check a page from one crawl had already been found in another. Second case may be if the data mining has to be perform upon the whole data set in an integrated way.
This architecture is a effective procedure to use idle computing resource within a industry. In order to produce the capacity to execute special types of large scale web data mining tasks.

Acknowledgement:
It gives me immense pleasure in Research on topic Search engine optimization-using crawler optimization. I acknowledge the enormous assistance and excellent co-operation extended by my respected Guide DR.S.K. Gupta ,Prof. , Department of Computer Sci. & Engg., BIET,Jhansi and presently Joint Controller and Examination, M.T.U., Noida.

References:

Authors:-

First Author - Rajesh Singh (Profile)
Received a B.E. in Computer Sci Engineering at Dr Bhim Rao Ambedkar University,Agra in 2000 and M.Tech in Information Technology from Karnataka State University, Karnataka He is currently a Ph.D. candidate in Computer Science at Bhagwant University, Ajmer (Raj.). Some of his research interests include the Data Mining, Computer Architecture, human computer interaction, search engines, scalability of information access interfaces.

Second Author -Dr. S. K. Gupta (Profile)