JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-5 E-ISSN: 2347-2693

Implementation of Mobile Optimized Search Crawler

Kukreja Kajal[1*], Gavali Nishigandha[2] and Khedlekar Gandhali[3]

[1*, 2, 3] *R. H. Sapat College of Engineering, Management Studies & Research, Nashik, Maharashtra, India*

Available online at: www.ijcseonline.org

Received: Apr/21/2016 Revised: May/04/2016 Accepted: May/18/2016 Published: May/31/2016

***Abstract* —** The web crawler is the central component of a search engine which works like an indexer, finds out hyperlinks, computes keyword density of each web page and stores the visited links for the future use. Today's world is mobile world! Aodhan Cullen, CEO, StatCounter stated that "Mobile has grown rapidly from 17.1% to 28.5% in past 12 months. Mobile usage has already overtaken desktop in several countries including India, South Africa and Saudi Arabia[1]". Due to increasing use of mobile, there is a need to develop a search crawler which will mainly focus on the mobile devices and provide them quality results in less time. The proposed system, "Mobile Optimized Search Crawler" is a crawler that mainly focuses on providing quick and relevant results to the mobile users. It gives more priority to the websites which are optimized for mobile devices than the websites which are not. It uses more than 50 factors to determine the relevancy of page, giving more weightage to the mobile-optimized factor so that mobile users get websites which are user-friendly and rich in knowledge.

***Keywords* —** Search engine; Search Crawler; Keyword density; Mobile-optimized website.

I. Introduction

A search engine is a system that searches for information on the World Wide Web and finds out the web pages that contain information related to the search keyword. "Search engines automatically create web site listings by using spiders that crawl the web pages, index their information, and optimally follows that site's links to other pages.[2]" A web crawler is a program that crawls web pages on the World Wide Web to read visible text, hyperlinks and content of the various tags used in the site, such as keyword rich meta tags. Using this information that is gathered by the crawler, a search engine determines what the site is about, computes its keyword density and ranks it according to keyword density and various other factors. There are more than 200 factors that are considered while determining the relevance of a page. Few such factors are content length, frequency of page updation, number of pages, mobile-optimized, social status, SSL certificate and there are many more.

Mobile Optimized Search Crawler is a crawler that mainly focuses on providing quick and relevant results to mobile users. It gives more priority to the websites which are optimized for mobile devices than the websites which are not. It uses more than 50 factors to determine the relevancy of page, giving more weightage to the mobile-optimized factor so that mobile users get websites which are user-friendly and rich in knowledge.[1]

This paper is organized in five sections as follows - Section 1 is Introduction, Section 2 discusses the literature survey done for the proposed system, Section 3 shows the working of proposed system, Section 4 describes the implementation details and technologies used for this system and Section 5 concludes with results and goals for future work.

II. LITERARURE SURVEY

Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang and Hai Jin proposed "SmartCrawler : A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces" which performed site-based searching for the center pages with the help of search engines, avoiding visiting a large number of pages. "To achieve more accurate results for a focused crawl, *SmartCrawler* ranks websites to prioritize highly relevant ones for a given topic. In the second stage, *SmartCrawler* achieves fast in-site searching by excavating most relevant links with an adaptive link-ranking.[3]" The only drawback of this system was that this crawler did not mainly focus on mobile devices.

Mehdi Bahrami, Mukesh Singhal and Zixuan Zhuang proposed "A Cloud-based Web Crawler Architecture" which used cloud computing features and the MapReduce programming technique to crawl the web. Crawling was done by distributed agents with each agent storing its own finding on a Cloud Azure Table (NoSQL database). "A cloud-based web crawler allows people to collect and mine web content without buying, installing and maintaining any infrastructure.[4]" This web crawler stored unstructured and massive amount of data on Azure Blob storage. They analyzed the performance and scalability of web crawler and described its advantages over traditional distributed web crawlers. Proposed system uses Spark which is 100 times faster than MapReduce.

[1] Contact Author - Kukreja Kajal (www.kajalkukreja.com)

Pavalam S. M., S. V. Kasmir Raja, Jawahar M. and Felix K. Akorli researched on “Web Crawler in Mobile Systems” and explained various concepts of web crawlers. “Web crawler is the central part of the search engine which browses through the hyperlinks and stores the visited links for the future use.[5]” They also explained the ways in which crawlers can be used in mobile systems and explored the different kinds of software used in mobile phones for crawling purposes. Thus, they identified the advantages of crawlers in mobile communications.

Vladislav Shkapenyuk and Torsten Suel proposed “Design and implementation of a high-performance distributed web crawler” which ran on network of workstations and crawled hundreds of web pages simultaneously for finding out relevant information. “The crawler scales to (at least) several hundred pages per second, is resilient against system crashes and other events, and can be adapted to various crawling applications.[6]”

Hardik P. Trivedi, Gaurav N. Daxini, Jignesh A. Oswal, Vinay D. Gor and Swati Mali presented “An Approach to Design Personalized Focused Crawler” in which they defined web page change detection policy for the design of a focused crawler. “The motivation behind developing a Personalized Focused Crawler is to provide targeted information to user i.e. providing information based on user’s interest solely.[7]”

III. WORKING

Mobile Optimized Search Crawler works in 2 stages. These stages are Index building and Searching.

1. **Index building -**

   1) Take seed URLs and search keyword as input
   2) Crawl seed URLs for specified keyword
   3) Calculate the keyword density based on keyword prominence and other factors and store it in database
   4) Parse the page to find out new links and add them to database for crawling in future

2. **Searching -**

   1) Take search keyword as input from user
   2) Check if result is present in previous searches
   3) If result is found in database, sort it based on keyword density and return as output
   4) If result is not found then show appropriate message to the user and initiate crawling process for that keyword in background so that it can be found in later searches

IV. IMPLEMENTATION

For designing a crawler that mainly focuses on mobile-optimized websites, we have used many factors for determining the relevancy of page. These factors have been suggested by “Google Ranking Factors: The Complete List[8]” . All the factors which have been implemented in proposed system are listed below -

1. Domain Age
2. Keyword Appears in Top Level Domain
3. Keyword As First Word in Domain
4. Domain registration length
5. Keyword in Subdomain Name
6. Exact Match Domain
7. Keyword in Title Tag
8. Title Tag Starts with Keyword
9. Keyword in Description Tag
10. Keyword Appears in H1 Tag
11. Keyword is Most Frequently Used Phrase in Document
12. Content Length
13. Keyword Density
14. Page Loading Speed via HTML
15. Duplicate Content
16. Rel=Canonical
17. Recency of Content Updates
18. Keyword Prominence
19. Keyword in H2, H3 Tags
20. Multimedia
21. Broken Links
22. HTML errors/W3C validation
23. URL Length
24. URL Path
25. Page Category
26. Keyword in URL
27. URL String
28. Bullets and Numbered Lists
29. Quantity of Other Keywords Page Ranks For
30. User Friendly Layout
31. Site Updates
32. Number of Pages
33. Presence of Sitemap
34. SSL Certificate
35. Mobile Optimized
36. YouTube
37. Site Usability
38. Social Shares of Referring Page
39. Keyword in Title
40. Organic Click Through Rate for a Keyword
41. Organic CTR for All Keywords
42. Repeat Traffic
43. Query Deserves Freshness
44. User Browsing History
45. User Search History
46. Number of Facebook Likes

47. Facebook Shares
48. Site Has Facebook Page and Likes

We have used most recent technologies to provide fast performance and high scalability to this crawler. Following technologies have been used to implement the Mobile Optimized Search Crawler -

- **Apache Hadoop -**

Apache Hadoop is an open source software framework that allows distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer for delivering highly-available service on top of a cluster of computers, each of which is prone to failures.

- **Apache HBase -**

Apache HBase is an open-source, distributed, versioned and non-relational(NoSQL) database that runs on top of Hadoop Distributed File System (HDFS). It is column-oriented key/value data store that provides real-time read/write access to large datasets. HBase can scale linearly to handle huge data sets with billions of rows and millions of columns and it can easily combine data sources that use a wide variety of different structures and schemas. It was modeled after “BigTable” - Google's proprietary NoSQL database. It provides BigTable-like capabilities for Hadoop.

- **Apache Spark -**

Apache Spark is an open source and fast engine for large-scale data processing. It can run programs up to 100x faster than Hadoop MapReduce in memory and 10x faster on disk. It can run on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3. It provides programmers with an application programming interface centered on a data structure called the Resilient Distributed Dataset (RDD), a read-only multiset of data items distributed over a cluster of machines that is maintained in a fault-tolerant way.

- **Java Servlets** -

Java Servlet is a Java object that responds to HTTP requests. It is part of a Java web application. It runs inside a Servlet container such as Apache Tomcat. A Servlet container may run multiple web applications at the same time, each having multiple servlets running inside. A Java web application can contain other components than servlets. It can also contain Java Server Pages (JSP), Java Server Faces (JSF) and Web Services.

- **Java Server Pages (JSP) -**

Java Server Page (JSP) is a technology for controlling the content or appearance of Web pages through the use of servlets, small programs that are specified in the Web page and run on the Web server to modify the Web page before it is sent to the user who requested it. JSP is similar to Microsoft's Active Server Page (ASP) technology. Java Server Page calls a Java program that is executed by the Web server whereas an Active Server Page contains a script that is interpreted by a Script interpreter (such as VBScript or JScript) before the page is sent to the user.

V. CONCLUSION AND FUTURE WORK

Due to increasing use of mobile devices, there was a need to develop a web crawler which will be specially optimized for mobile devices. Thus, web crawler had a big scope for mobile systems. Looking at this scope, we have proposed a crawler which mainly focuses on websites which are optimized for mobiles. This crawler will eliminate the websites which are not user-friendly and mobile-friendly and obtain accurate search results for the mobile users in optimized time. The proposed system uses Apache Hadoop so in future, we plan to move on cloud to provide more storage and high performance. We also plan to integrate user security measures by removing spam websites from search results.

ACKNOWLEDGMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without mentioning the people who made it possible. We are grateful to a number of individuals whose professional guidance along with encouragement have made it very pleasant endeavor to complete this project.

We have a great pleasure in presenting “Implementation of Mobile Optimized Search Crawler” under the guidance of Prof. Dr. S. V. Gumaste and our project coordinator Prof. C. R. Barde. We are truly indebted and grateful to Head of Computer Engineering Department Prof. N. V. Alone for his valuable guidance and encouragement.

We would also like to thank Gokhale Education Society’s R. H. Sapat College of Engineering, Management Studies and Research, Nashik-5 for providing the required facilities, internet access and important books. At last, we must express our heartfelt gratitude to all the teaching and nonteaching staff members of Computer Engineering Department who helped us by giving their valuable time, support, comments and suggestions.

REFERENCES

[1] "StatCounter", http://gs.statcounter.com/press/mobile-internet-usage-soars-by-67-perc , 19 Dec, 2015.

[2] "Search Engine", http://websearch.about.com/od/enginesanddirectories/a/searchengine.htm , 15 July, 2015.

[3] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang and Hai Jin, "SmartCrawler : A two-stage crawler for efficiently harvesting deep-web interfaces", Services Computing, IEEE Transactions, Volume-**PP**, Issue-**99**, DOI-**10.11.09**, **2015**.

[4] Mehdi Bahrami, Mukesh Singhal and Zixuan Zhuang, "A Cloud-based Web Crawler Architecture", Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference, Page No (**216-233**), 17-19 Feb **2015**.

[5] Pavalam S.M, S.V. KumarRaja, M. Jawhar and Felix K. Akorli, "Web crawler in mobile systems", International Journal of Machine Learning and Computing, Volume-**02**, Issue-**04**, Page No (**531-534**), August **2012**.

[6] Vladislav Shkapenyuk and Torsten Suel, "Design and implementation of a high-performance distributed Web crawler", Data Engineering, 2002. IEEE Proceedings. 18th International Conference, Page No (**357-368**), **2002**.

[7] Hardik P. Trivedi, Gaurav N. Daxini, Jignesh A. Oswal, Vinay D. Gor and Swati Mali, "An Approach to Design Personalized Focused Crawler", International Journal of Computer Sciences and Engineering, Volume-**02**, Issue-**03**, Page No (**144-147**), March **2014**.

[8] "Google Ranking Factors: The Complete List", http://www.backlinko.com/google-ranking-factors , 2 Feb, 2016.

[9] "Apache Hadoop", http://www.hadoop.apache.org , 10 March, 2016.

[10] "Apache HBase", http://www.hbase.apache.org , 10 March, 2016.

[11] "Apache Spark", http://www.spark.apache.org , 14 March, 2016.

[12] "Java Servlets", http://docs.oracle.com/javaee/5/tutorial/doc/bnafd.html , 14 March, 2016.

[13] "Java Server Pages", http://docs.oracle.com/javaee/5/tutorial/doc/bnagx.html , 15 March, 2016.

AUTHORS' PROFILE

Miss Kajal Kukreja has completed Diploma in Information Technology with distinction from Government Polytechnic, Nashik. She is pursuing Bachelors degree in Computer Engineering from R. H. Sapat College of Engineering, Management Studies & Research, Nashik. After working as a web developer for three months in a company, she is currently working as a freelancer and has deployed more than 65 projects till date. She is a passionate web developer and software programmer with an ability to convert client requirements into innovative solutions. She has designed her professional website with URL http://www.kajalkukreja.com

Miss Nishigandha Gavali has completed Diploma in Computer Technology from Government Polytechnic, Nashik. She is pursuing Bachelors degree in Computer Engineering from R. H. Sapat College of Engineering, Management Studies & Research, Nashik. She looks forward to have a successful career in Graphic Designing, Virtualization, Cloud Computing and Distributed Computing.

Miss Gandhali Khedlekar is pursuing Bachelors degree in Computer Engineering from R. H. Sapat College of Engineering, Management Studies & Research, Nashik. She has completed HSC (12th) from R. Y. K. College of Science, Nashik. Her areas of interest include Mobile Computing and Cloud Computing.