Web Crawler in Mobile Systems

Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli

Abstract—With the advent of internet technology, data has exploded to a considerable amount. Large volumes of data can be explored easily through search engines, to extract valuable information. Web crawlers are an indispensable part of search engine, which are program (proceeds with the search term) that can traverse through the hyperlinks, indexes them, parses the files and adds new links in to its queue and the mentioned process is done several times until search term vanishes from those pages. The web crawler looks for updating the links which has already been indexed. This paper briefly reviews the concepts of web crawler, its architecture and its different types. It lists the software used by various mobile systems and also explores the ways of usage of web crawler in mobile systems and reveals the possibility for further research.

Index Term—Web crawlers, mobile systems, mobile web crawler.

I. INTRODUCTION

World shrinks in to a tiny mass through mobile phones, whereby communications has been made at ease, searching relevant things in a moment, acting as a device for location, marketing tools, etc… Thanks to the advent of internet, web search engine helps us to search relevant information from enormous volumes of data without wasting time, effort and looking for physical resources. Studies show that mobile access to internet exceeded the desktop access to internet [1]. Mobile has seamlessly integrated in to everybody’s life.

Web crawler is the central part of the search engine which browses through the hyperlinks and stores the visited links for the future use. Different search engines indexes using different techniques [2]. Crawler looks for the modifications for the indexed pages by revisiting them and updates its store by deleting the previous [3]. The crawler is intended to traverse to huge number of pages to maintain with enormous growth of data, to improve the efficiency and effectiveness of search, to retrieve relevant information and be user friendly. Mobile systems can’t access many pages that are accessible in desktop and also speed is considerably low [4]. So there is a strong need for optimized browsers suitable for quick information access in mobile systems. Literature reveals that this research area has scope for more exploration.

This paper explores the concepts of web crawler and ways of its usage in mobile systems. This work is organized as follows. Section 1 introduces the web crawler, section 2 discusses the architecture of the crawler and the working of the crawler, Section 3 discusses the different types of crawlers, and Section 4 explores the software used in mobile phones for crawling purposes with Section 5 discusses the advantages that crawlers can bring in mobile communications and section 6 brings out the summary.

II. ARCHITECTURE OF A CRAWLER

Web crawler (also known as Spiders or robots) is software that can start with a Uniform Resource Locator (known as seed URL), downloads the pages with associated links and looks for the updates and stores them for later use. This process is done iteratively. The general architecture of a crawler is portrayed in Fig. 1.

HTTP (HyperText Transfer Protocol) request is being sent to the World Wide Web to download the pages by giving a seed URL. The pages are retrieved by the web crawler and follow the link available on that page. It is sent to the parser – a major component in the crawling technology, which actually checks whether relevant information is retrieved [5]. The relevant contents are then indexed by the indexer [6] and it is stored for later use. The crawler looks for the updates on the indexed pages and if so the old information is replaced with the new information, until the search term vanishes from those pages.

When a user enters a search engine page (Input or Seed URL) and places a keyword, the crawler visit the links and request a file called “robots. txt” (a file which defines the limitations – what it can let it to see, how many times it can allows visits, etc )gets the copies and it will be consulted with the indexer, and the relevant pages are given out. The results are ranked and best relevant results are leading the priority in the display. The methods of ranking and the order of display vary from one search engine to another [7]. The crawler is smart enough to check out the pages after some time, when the site is down temporarily, but if it finds the site down for continuous period or too slow to respond it may not prefer to visit again.

III. TYPES OF CRAWLER

Different strategies are being employed in web crawling. These are as follows.
A. Focused Web Crawler

A focused web crawler returns pages which are specific and relevant to the given topic. The focused crawler determines the following – Relevancy, Way forward. It determines how far the given page is relevant to the particular topic and how to proceed forward [8]. The advantage of focused web crawler is that it is economically feasible in terms of hardware and network resources and also its search exposure is huge [9]. It employs different techniques for searching. Certain focused crawlers employ best fit search strategy. Some focused crawlers employ page rank technique for giving out the most important page [10]. Others uses neural net, back propagation to find the most relevant [11]. Several other techniques are being used for focused crawling [12].

B. Incremental Crawler

An incremental crawler is one which updates its index collection on an incremental basis after its target accumulation is finally reached and based on an estimate [13]. It refreshes the existing collection by new updates on a periodical basis [14]. It helps to save network bandwidth and also effective [15]. Several approaches [16], [17], [18], [19] has been used.

C. Distributed Crawler

Many crawlers are employed to distribute the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed [20]. It needs increased computer nodes and storage capability to increase crawling efficiency [21]. It basically uses Page rank algorithm for its increased efficiency and quality search [22]. There have been many other approaches [23], [24], [25], [26] which are proposed.

D. Parallel Crawler

Multiple crawlers are often run in parallel, which are referred as Parallel crawlers [27]. The Parallel crawlers depend on Page freshness and Page Selection [28]. A Parallel crawler can be on local network or be distributed at geographically distant locations [29], [30], [27], [31] have proposed interesting and different methods for achieving high performance and effective memory usage.

IV. SOFTWARE USED IN WEB CRAWLING

In this section we intend to present the software which is used in different mobile systems. The basic underlying technologies are HTML, XHTML, WAP, WML, CSS, ECMA Script.

<table>
<thead>
<tr>
<th>Browser</th>
<th>Origin</th>
<th>Used in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polaris Browser</td>
<td>Infraware Inc.</td>
<td>Nokia, Samsung, LG Electronics, KYOCERA and other Smartphone and cellular phone in USA, China, Korea, etc</td>
</tr>
<tr>
<td>Kindle Basic Web</td>
<td>Amazon.com</td>
<td>Blackberry</td>
</tr>
</tbody>
</table>

Table I lists the compilation of various browsers in use by different Mobile systems from the web.

V. APPLICATIONS OF CRAWLER IN MOBILE SYSTEMS

A. Mobile Learning

eLearning is a technology, which helps learning to take place at user paced and also fitting their interest and level of understanding [32]. In eLearning Learners are exposed with a wide variety of resources for education, whereby the relevant material is found by crawling through the web [33]. It also enhances in Human – computer interaction [34]. Web mining techniques like web logs, web usage mining could help to extract the interests of the learners and help to improve the learning environment. The Mobile Learning is an extended concept of eLearning through mobile means. Research [35] shows mobile could be an excellent tool in instigating learning process. Digital libraries act as a vital tool in the learning process.

B. Mobile Commerce

Mobile commerce is the application of wireless technology to the industry of commerce. Mobile commerce applications are having a wide coverage. Some of them are mobile advertisement, mobile banking, mobile location finder,
mobile shopping [36]. Mobile advertisement serves as means of advertisement targeted to the people on specific location through mobile devices (for example information about offers sale for products pertained to a specific location people). Mobile banking is the most popular and economical means of transaction services through the mobile phones which helps to improve the quality of services and business processes. Mobile serves as a geographic location finder which adds significant importance by giving information available in that particular area. Shopping is an integral part of human life and mobile adds significant value by giving the information for the queried products as well as relevant products, which helps saving time and energy. [37] Shows the disposition of people on this mobile technology.

C. Social Relations

Mobile plays a major role in establishing and maintaining social relationships. Social networking helps the users to connect virtually at all times, leaving behind the feeling of disconnected from the social activities of life, even in their busy lifestyle. It covers a wider range of applications in mobile - simple text messaging, mobile communication, connecting to a large network of specific interests (ex: facebook), active email alerts, analyzing the potential customers of the future and E-Governance [38], [39].

VI. SUMMARY

This paper gave a snapshot of Web Crawling aspects, its architecture and its techniques. It also explored the different software used in browsing in mobile. The overview presented in this paper shows web crawler has a significant scope on the Mobile systems. We then proceeded to explore some potential applications of web crawler in the Learning field, commercial field and Social relationship. In closing the scope of Web Crawler in Mobile needs to be explored further.

REFERENCES


Pavalam S. M. is currently working at National University of Rwanda and also doing her Ph.D in Computer Science and Engineering at SRM University, INDIA. Her research interest includes Data Mining, Information Retrieval, Web crawling, Mobile Technology, E-Learning/M-Learning. She has many International Publications in Journals and conferences.

Dr. S. V. Kashmir Raja is currently the Dean Research at SRM University, India. He received his Ph.D from Indian Institute of Science, Bangalore in 1986. His research interest includes computer networks and software engineering. He has more than 60 publications in International/National Journals and conferences. He has guided several Ph.D students. He has received several awards such as “The Man of Year – 1999” by American Bibliographical Institute - USA, and “2000 Outstanding Scholar of 20th Century” by International Bibliographical Centre, Cambridge – England.

Jawahar M. is currently the Director of ICT at the National University of Rwanda. His research interest includes E-Learning/M-Learning, Networking. He has keen interest in Software development and Implementation. He has many International Publications in Journals and conferences.

Dr. Felix K. Akorli is currently coordinator of M.Sc ICT at National University of Rwanda. He has many International Publications in Journals and conferences. His research interests include Electronics, Mobile communications.