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 indispensible 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.

Manuscript received May 30, 2012; revised July 16, 2012.

Pavalam is with the National University of Rwanda, Rwanda, India (e-mail: spavalam@nur.ac.rw).

Kasmir Raja is with SRM University, Chennai, India. (e-mail: svkr@srmuniv.ac.in).

Jawahar M. and Felix Akorli are with ICT at National University of Rwanda, Rwanda, India. (e-mail: mjawahar@nur.ac.rw; fakorli@nur.ac.rw).

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.

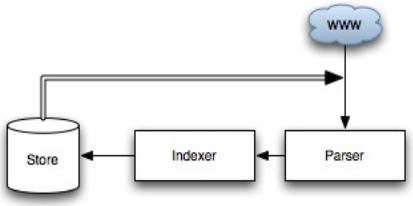

Fig 1. Architecture of a web crawler

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 the different techniques for searching. Certain focused crawlers employ Best – Fit search strategy. Some focused crawler employs 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 updations 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 in 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 I: Browsers in Various Mobile Systems

Browser	Origin	Used in
Polaris Browser	Infraware Inc.	Nokia, Samsung, LG Electronics, KYOCERA and other Smartphone and cellular phone in USA, China, Korea, etc
Kindle Basic Web	Amazon.com	Blackberry
Android browser	Google	Nexus
WebOS Browser	Palm	-
BlackBerry Browser	Research in Motion	-
Blazer	Palm	installed on all newer Palm Treos and PDAs
Firefox for mobile	Mozilla	Nokia Maemo and for Android
Internet Explorer Mobile	Microsoft	Samsung newer versions
Iris Browser	Torch Mobile Inc.	Acquired by Research in Motion - No longer supports Windows Mobile or Linux
Myriad Browser (Previously Openwave Mobile Browser)	Myriad Group	Acquired from Openwave in 2008
NetFront	ACCESS Co., Ltd.	-
Nokia Series 40 Browser	Nokia	-
Obigo Browser	Obigo AB	100% owned by Teleca AB
Opera Mobile	Opera Software	Capable of reading HTML and reformat for small screens, installed on many phones
PlayStation Portable web browser	Sony	Sony PSP
Safari	Apple Inc	on iPhone, iPod Touch and iPad
Skyfire Mobile Browser	Skyfire	Renders Flash 10, Ajax and Silverlight content. supports Windows Mobile 5/6.x,Symbian S60 and Android.
uZard Web	Logicplant Co., Ltd.	on Samsung, LG Electronics and other smartphones and cellular phones in Korea
xScope	xScope Mobile	on android systems

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

[1] International Telecommunications union, "The world in 2009: ICT facts and figures" presented at ITU Telecom world2009 Geneva, Oct. 2009.

[2] Stefan Buttcher, Charles L. A. Clarke, and Gordon V Cormack, "*Information Retrieval*" MIT Press, 2010, ch I.2, pp. 40-50.

[3] Satinder Bal and Rajender Nath, "A Novel Approach to Filter Non-Modified pages at remote site without downloading during crawling" *Proc. International Conference on Advances in Recent Technologies in Communication and Computing 2009 (ARTCom 2009), IEEE*, pp. 165 – 169.

[4] Arie van der Dussen, Elizabeth Burd "*Reverse Engineering"* IEEE Computer society, 2002, ch 1, pp14-16.

[5] N. A. El-Ramly, H. M. Harb, M. Amin, and A. M.Tolba, "More Effective, Efficient, Scalable Web Crawler Architecture" *Proc of Electrical, Electronic and Computer Engineering, 2004. ICEEC '04*, pp. 120-123.

[6] Muhammad shoaib and Shazia Arshad, "Design and implementation of web information gathering system" *Second National Information Technology Symposium (NITS 2007), CCIS*, paper 142 – 1.

[7] Mohammed Khan "*Search Engine Optimization – what do you need to know*" SEO expert, 2008, pp. 22.

[8] AH Chung Tsol, Daniele Forsali, Marco Gori, Markus Hagenbuchner, Franco Scarselli, "A Simple Focused Crawler" *Proc 12th International WWW Conference 2003(poster)*, pp. 1.

[9] Soumen Chakrabarti, Martin van den Berg, and Byron Dom "Focused crawling: a new approach to topic-specific Web resource discovery" *Proc of 8th international WWW conference (WWW8)*, Toronto '99, pp. 545 – 562.

[10] Junghoo Cho, Hector Garcia-Molina, Lawrence Page "Efficient Crawling Through URL Ordering*", Proc of the 7th International WWW Conference*, 1998, pp. 161-172.

[11] Gatial E., Balogh Z., Laclavik M., Ciglan M. and Hluchy L. " Focused Web CrawlingMechanism based on Page Relevance*" Proc ITAT 2005 Information Technologies - Applications and Theory, Peter Vojtas (Ed.)*, pp.41-46.

[12] Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang, "A Survey in Semantic Web Technologies-Inspired Focused Crawlers", *Proc ICDIM 2008*, pp. 934 – 936.

[13] A. K. Sharma and Ashutosh Dixit, "Self Adjusting Refresh Time Based Architecture for Incremental Web Crawler" *International Journal of Computer Science and Network Security*, vol.8 no.12, 2008, pp. 349-354

[14] Niraj Singhal, Ashutosh Dixit, and Dr. A. K. Sharma "Design of a Priority Based Frequency Regulated Incremental Crawler" *International Journal of Computer Applications* (0975 – 8887) vol.1, no. 1, 2010, pp. 42-47.

[15] Junghoo Cho and Hector Garcia-Molina, "The Evolution of the Web and Implications for an incremental Crawler" *Proc VLDB Conference*, 2000, pp. 200 – 209.

[16] Qingzhao Tan and Prasenjit Mitra, "Clustering-based IncrementalWeb Crawling" *ACM Transactions on Information Systems*, vol. 2, no. 3, 03 2010, pp. 1–25.

[17] Wei Liu, Jianguo Xiao, and Jianwu Yang, "A Sample-Guided Approach to Incremental Structured Web Database Crawling" *Proc IEEE International Conference on Information and Automation 2010*, IEEE, pp. 890-895

[18] Christos Bouras, Vassilis Poulopoulos, and Athena Thanou " Creating A Polite, Adaptive And Selective Incremental Crawler" *Proc International Conference WWW/INTERNET 2005, Portugal*, vol. I, pp.307-314.

[19] Jiang-Ming Yang, Rui Cai, Chunsong Wang, Hua Huang, Lei Zhang, and Wei-Ying Ma "Incorporating Site-Level Knowledge for Incremental Crawling of Web Forums: A List-wise Strategy" *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining KDD'09*, 2009, Paris, pp1375-1384

[20] Kunpeng Zhu, Zhiming Xu, Xiaolong Wang, and Yuming Zhao, "A Full Distributed Web Crawler Based on Structured Network": *AIRS 2008, Lecture Notes in Computer Science* , H. Li et al. (Eds.), pp. 478–483

[21] Jos´e E x p o s to, Joaquim Macedo, A n t´onio Pina, Albano Alves,and Jos´e Rufino, "Efficient Partitioning Strategies for Distributed Web Crawling" ICOIN 2007, 2008, *Lecture Notes in Computer Science* , T. Vaz˜ao, M.M. Freire, and I. Chong (Eds.), pp. 544–553.

[22] S. Brin and L. Page " The Anatomy of a Large Scale Hypertextual web search engine" *Proc. 7th International world wide web conference*, 1998, pp. 107-117.

[23] Marc Najork, Allan Heydon "*Handbook of Massive Data Sets*" (edited by J. Abello, P. Pardalos, and M. Resende), Kluwer Academic Publishers, Inc., 2001, ch2, pp. 28-35.

[24] Apostolos Kritikopoulos, Martha Sideri, and Kostantinos Stroggilos, "Crawl Wave: A Distributed Crawler," *Proc 3rd Hellenic Conference on Artificial IntelligenceSETN04*, 2004, *LNAI*.

[25] Kunpeng Zhu, Zhiming Xu, Xiaolong Wang, and Yuming Zhao "A Full Distributed Web Crawler Based On Structured Network" *Lecture Notes in Computer Science* 2008, vol 4993/2008 pp. 478-483.

[26] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna, "UbiCrawler: a scalable fully distributed web crawler", *Software—Practice & Experience*, 2004, vol.34 no.8, pp.711-726.

[27] Shoubin Dong, Xiaofeng Lu, Ling Zhang and Kejing He "An Efficient Parallel Crawler In Grid Environment" *Lecture Notes in Computer Science* 2004, vol. 3032/2004 pp. 229-232.

[28] Junghoo Cho and Hector Garcia Molina "Parallel Crawlers" *Proc. 11th international conference on world wide web WWW2002*, pp.124-135.

[29] Nidhi Tyagi and Deepti Gupta, "A Novel Architecture for Domain Specific Parallel Crawler" *Indian Journal of Computer Science and Engineering*, vol. 1 no. 1 44-53, pp. 44-53.

[30] Divakar Yadav, AK Sharma, and J. P. Gupta "Parallel Crawler Architecture and Web Page Change Detection", *WSEAS Transactions On Computers*, Issue 7, Volume 7, July 2008, pp929-940.

[31] Mauricion Marin, Rodrigo Paredes, and Carolina Bonacic "High Performance Priority Queues For Parallel Crawlers" *Proc. 10th ACM Workshop On Web Information And Data Management (WIDM'08)*, pp. 47-54.

[32] Yevgen Biletskiy, Michael Wojcenovic, and Hamidreza Baghi, "Focused Crawling for Downloading Learning Objects – An Architectural Perspective" *Interdisciplinary Journal of E-Learning and Learning Objects*, Volume 5, 2009, pp. 169 – 180.

[33] Julien Tane, Christoph Schmitz and Gerd Stumme, "Semantic Resource Management for the Web:An ELearning Application" *Proc WWW2004*, pp. 1-10.
[34] O. Shata, "A Personalized Multimedia Web-Based Educational System with Automatic Indexing for Multimedia Courses" *Proc World Congress on Engineering 2007 (WCE 2007)*, vol. I, pp. 269 – 273.
[35] Kumar, L. S., Jamatia, B., and Aggarwal A. K., "Mobile phones as effective learner support devices: A case study" *Proc International Conference on Distance and Learning Education (ICDLE 2010)*, pp. 211- 213.
[36] Hao Huang, Lu Liu, and Jianjun Wang, "Diffusion of Mobile Commerce Application in the Market" *Proc. Second International Conference on Innovative Computing, Information and Control, 2007. (ICICIC '07*), pp. 485-488
[37] Manochehri, N.-N. and AlHinai, Y. S.,"Mobile-phone users' attitudes towards' mobile commerce & services in the Gulf Cooperation Council countries: Case study" *Proc International Conference on Service Systems and Service Management*, 2008 , pp. 1-6.
[38] Georgios Lappas, "An Overview of Web Mining in Societal Benefit Areas" *Proc The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007)*, pp. 683-690.
[39] Scott Counts, Karen E. Fisher "Mobile Social Networking: An Information Grounds Perspective" *Proc 41st Hawaii International Conference on System Sciences* – 2008, pp. 1-10.

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.