A REVIEW OF WEB CRAWLERS FOR INFORMATION RETRIEVAL

[1]Babatunde Ajose-Ismail & [2]Ayodeji Osanyin
[1]Department of Computer Science, School of Applied Science
The Federal Polytechnic Ilaro, Ogun State. Nigeria.
babatunde.ajose@federalpolyilaro.edu.ng
[2]Department of Computer Science, School of Applied Science,
Lm Ericsson Nig Limited .
Osanyindeji@gmail.com

Abstract

Performance of any search engine relies heavily on its Web crawler. Web crawlers are the programs that get webpages from the web by following hyperlinks. These webpages are indexed by a search engine and can be retrieved by a user query. In the area of web crawling which is a subfield of Information Retrieval, we still lack an exhaustive study that covers all crawling techniques. This study follows the guidelines of systematic literature review and applies it to the field of Web crawling. Existing literature about the web crawler is classified into different key subareas. Each subarea is further divided according to the techniques being used. We have highlighted future areas of research. We call for an increased awareness in various fields of the web crawler.

Keywords: Web Crawler, Information Retrieval, Search Engine, Webpages, User Query

Introduction

The profound depth of the web is so intense and filled with vast amount of webpages that finding information is like finding a needle in a haystack. To ease the process of locating vital information on the web, different search engines have been developed by top companies like Google, Yahoo, Bing and Ask amongst others. This search engines have become an essential part of our digital life.

Kumar, Bhatia, and Rattan, (2017) opined that search engines fall under two categories- crawler based and human powered. The human powered search engine indexes a collection of high-quality user submitted or handpicked websites. Human Intervention often affects the results or position of results in human powered search engines. Crawler-based search engine have three main components: crawler, indexer, and searching-ranking algorithm

A web crawler is the heart of any crawler-based search engine. Technically, web crawlers are the tools for data acquisition in the search engines. They are also called spiders or robots or wanderers. To crawl means to move in one direction slowly and a crawler works by traversing webpages on the web to gather information that can be indexed by the indexer to handle any user query efficiently. Web crawler starts the process of crawling from a single uniform resource locator (URL) or a set of seed URLs. As a crawler visits a URL, it adds all hyperlinks in the webpage to a list of URLs to be visited further. The objective of crawling is to collect as many useful webpages as possible in the least possible time

Searching ranking algorithm returns those webpages that are best matched and ordered in response to respective user query. Mostly a user analyzes first few results in response to the query he has submitted. So, it becomes necessary to order the results efficiently.

The motivation behind our study is to explore various strategies used for web crawling.

The remainder of this paper is organized as follows: Background section discusses the architecture, types, policies, and challenges in the area of a Web crawler.

Current status of web crawler section is also discussed. In the next section, various performance metrics for web crawler are presented. The fifth section discusses various future avenues for carrying out research in the area of web crawlers. Final section concludes and provides recommendations for future work

Background

In this section, we discuss various terminology related to the web crawler. We present a taxonomy of web crawlers and summarize different types of crawler.

Architecture of a Web Crawler

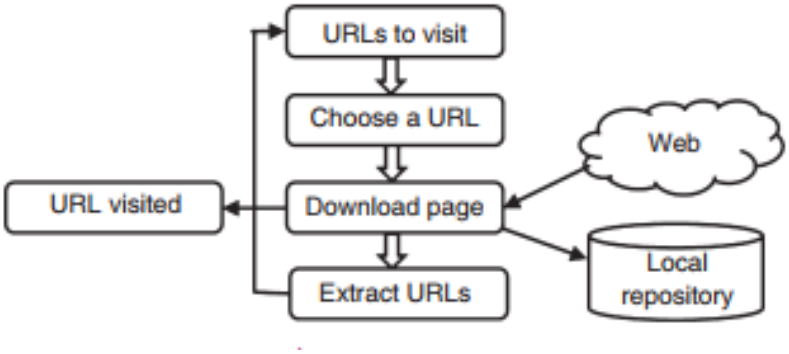

Figure 1: Architecture of Web Crawler (Source: Kumar, Bhatia, & Rattan, 2017)

Types of Web Crawler

Various researchers have categorized web crawler in different way. Figure 2 gives a broad taxonomy of web crawler

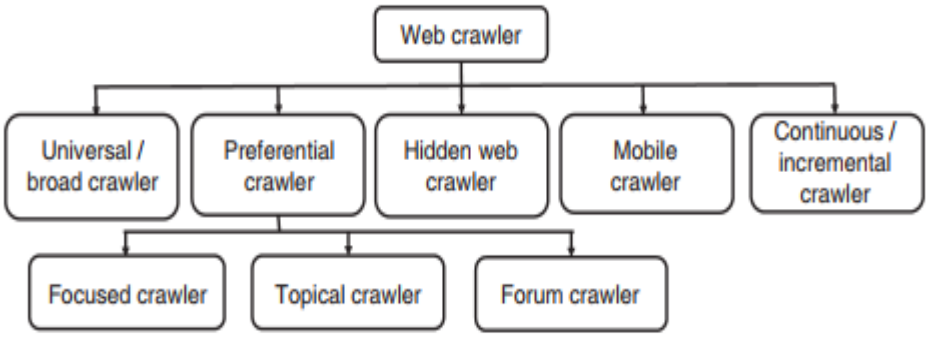

Figure 2: A taxonomy of Web Crawler (Source: Kumar et al., 2017)

Universal or Broad crawler

These type of web crawlers are not limited to webpages of a particular topic or domain. They keep on following links endlessly and get all webpages they encounter.

Preferential crawler

This category of web crawler does not crawl all links they encounter rather the user submits a condition or topic of interest that guides the preferential crawler. Furthermore, the preferential crawler can be categorized as focused and topical crawler. Chakrabarti, Van Den Berg, and Dom (1999) proposed one of the first focused crawler that selectively seeks out webpages that are relevant to a predefined set of topics

Preferential crawler

Topical crawler or topic-specific crawler is used for searching information related to some specific topic from the web. Topical crawling assumes that only the topic of interest is specified while focused crawling assumes that some labeled examples of relevant and non-relevant webpages are also available (Yu, Bingwu, & Fang, 2010). The other category of forum crawler only deals with crawling of online forum content

Hidden Web crawler

A significant amount of information on the web cannot be accessed directly by following the hyperlinks on webpages. This information is hidden behind search or query interface, this part of the Web is called hidden web r deep Web (Gravano, Ipeirotis, Sahami, & Prober 2003).A special category of crawlers called hidden Web crawler deals with crawling this section of the Web

Mobile crawler:

It is a method of crawling in which selection and filtration of webpages are performed on server side rather than on the search engine side. Moving code to data in mobile crawling reduces network load caused by traditional web crawler (Hammer & Fiedler, 2000)

Continuous or Incremental crawler

The web is dynamic and data on the webpages keep on changing frequently. These crawlers are used to maintain the index database of the search engine up-to-date (Badawi, Mohamed, Hussein, & Gheith, 2013).However, there is a trade-off between managing freshness and resources consumption

Web Crawling Policies

Any web crawler has the potential to disrupt the services of a server. Koster (1994) gives a set of policies every web crawler must follow

***Politeness policy*:** A crawler should not hamper any website with the requests. Every crawler is expected to respect Robots.txt and should crawl only allowed webpages. Crawler should act as a 'good citizen' of the Web world.

***Parallelization policy*:** Multiple threads of a crawler are used to maximize download rate of webpages. A policy is required for assigning new URLs discovered during crawling process to different threads running in parallel.

***Revisit policy*:** To keep an index of a search engine up-to-date, a Web crawler needs to revisit webpages. Either a uniform revisit policy or a proportional revisit policy can be used for deciding when to revisit a webpage.

***Robustness policy*:** The web is hosted by servers that may mislead a crawler and get it stuck into fetching a huge number of webpages of a particular domain. However, not all the traps are malicious some are due to side effects of faulty website designs. A crawler must be immune to malicious behavior of any web server.

Challenges in Crawling the Web

Regardless of the category of a web crawler, the researchers when dealing with web crawlers face some challenges. Some of the challenges are listed here.

***Non-uniform structures*:** The web is dynamic and uses inconsistent data structures, as there is no universal norm to build a website. Due to lack of uniformity, collecting data becomes difficult. The problem is amplified when crawler has to deal with semi-structured and unstructured data. (Abiteboul, 1997)

***Scale and revisit*:** Size of the web cannot be measured. Furthermore, there is a trade-off between coverage and maintaining freshness of a search engine database. The goal of any web crawler must be to ensure coverage of all web crawler reasonable content while bypassing low quality and irrelevant content (Olston & Najork, 2010)

***Crawling multimedia*:** A crawler can easily analyze text but analyzing multimedia is an open challenge. Analyzing multimedia content on webpages to detect criminal activities is one of the prominent applications these days. (Turek, Opalinski, & kisiel-Dorohinicki, 2011)

***Crawling deep Web*:** A large part of the web is hidden behind search interfaces and forms. This part of the web that cannot be reached directly comprises hidden web or deep web. Hidden web is accessible by querying the database, but query selection is another challenge (Zheng, Wu, Cheng, & Jiang, 2013)

Current Status of Web Crawlers

Focused Crawler Techniques

Focused crawler technique gives priority to those URLs in the process of crawling, in which probability of finding information of user's interest is high. Techniques used by the focused crawler, topical crawler, and forum crawler are somewhat the same. Most of the studies present in literature are related to the focused crawler (Dong & Hussain, 2013)

Hidden Crawler Techniques

The part of web that is beyond login form, search, or query interfaces is called hidden web. A large amount of information on the Web is hidden behind search interfaces and forms that cannot be indexed by a traditional web crawler. Crawling hidden web is an open challenge due to its heterogeneous and dynamic nature (Singh & Sharma, 2013)

Mobile Crawler Techniques

A mobile crawler is a program that can transfer itself to the Web server to download information and contents available on the server. The mobile crawler uses the approach of bringing 'code to data' instead of traditional 'data to code' approach. There are few studies about the mobile crawlers that are available in the literature that are categorized as shown in Table 1.

Table 1: Techniques used by Focused, Hidden and Web Crawler

S/N	Category	Techniques
1	Focused Crawler	Soft Computing Techniques (Ning, Wu, He, & Tan, 2011; Dong & Hussain, 2013; Du, Hai, Xie, & Wang, 2014), Link, Text and URL Technique (Ahmadi-Abkenari, & Selamat, 2012; Bošnjak et al., 2012) Parallel and distributed Technique (Ahmadi-Abkenari, & Selamat, 2012; Yani & Wibowo, 2013), Incremental and Revisit Policy (Mali & Meshram, 2011)
2	Hidden Crawler	Form-Based Approach (Nguyen, Nguyen, & Freire, 2010; Furche, 2013), Revisit policy and Incremental approach (Madaan, Dixit, Sharma, & Bhatia, 2010; Zhang, Dong, Peng, & Yan, 2011)
3	Mobile Crawler	Agent-based mobile crawler(Brin & Page, 2012), Freshness and revisit policy based mobile crawler (Rattan, Bhatia, & Singh, 2013)

Performance Metrics for Web Crawlers

A Web crawler performance metric is a standard measure of a degree to which a crawler possesses some property. The following are some performance metrics used by crawlers

Precision: precision as the proportion of retrieved material that is actually relevant

Recall: recall as the proportion of relevant material retrieved in answer to a search request

***F1-measure*:** that is the harmonic mean of precision and recall. When the value of harmonic mean reaches the highest, it signify the values of precision and recall reaches to highest at the same time

Harvest Ratio: is defined as the fraction of webpages crawled that satisfy the relevance criteria among all crawled webpages

Accuracy Ratio: is defined as the ratio of detected webpages that are related to the topic to the sum of detected webpages by the crawler

However, it is worthy of note that the performance metrics used by each of the crawlers vary but precision and recall are the most commonly used and fairly understood quantities in the field of information retrieval. Also, evaluation of many studies in information retrieval is done by using precision and recall for two or more crawlers on the same set of data.

Table 2: Performance Metrics Used by Various Studies

S/N	Category	Citation
1	Precision (Relevance)	Abbasi, Fu, Zeng, & Adjeroh, 2013; Ahmadi-Abkenari, & Selamat, 2012; Baykan, Henzinger & Weber, 2013; Chen, Liu , Zhai , Jiang & Cao, 2012; Chy, Seddiqui, & Das, 2014; Yani & Wibowo, 2013
2	Recall (Coverage)	Abbasi et al., 2013; Ahmadi-Abkenari, & Selamat, 2012; Baykan et al, 2013; Dong, 2014; Uzun, 2014
3	Harvest Ratio	Agarwal & Sureka, 2014; Zheng et al., 2013
4	F1-Measure	Abbasi, Fu, Zeng, & Adjeroh, 2013; Baykan, Henzinger & Weber, 2013
5	Accuracy	Ning, Wu, He, & Tan., 2011; Uzun, 2014

Discussion

This work can benefit the researchers who are looking for open research issues in the area of web crawler and for the practitioner to determine the best technique as per their needs. The crawler technique to be used depends on various factors like the resources available to the web crawler, the amount of relevant information about the topic of interest on the web, the total number of websites to be crawled, the rate of change of the webpages, and amount of malicious content on the webpages. The area of focused crawler is evolving at a rapid speed. Mobile crawler is still growing and can be the future of crawling. Less work has been done in this area and there are many possibilities in the field of mobile crawler. We consider the aforementioned avenues as the most promising for future research.

Conclusion

Extracted studies were categorized into different areas of a web crawler. A detailed description of each category is presented. Our review shows that researchers are working intensively in the field of web crawler. We observed that web crawler is used extensively for applications that target a particular group of users. A web crawler that deals with getting data from hidden web is also of interest for the researchers around the globe. To handle the exponential growth of the web, mobile crawler will get popularity as times goes on. We believe that this work will act as a research ground for the key areas of web crawlers. This study will be beneficial not only to researchers but also to practitioner and developers to get an insight of web crawlers.

References

Abbasi, A., Fu, T., Zeng, D., & Adjeroh, D. (2013). Crawling credible online medical sentiments for social intelligence. In 2013 International Conference on Social Computing (pp. 254-263). IEEE.

Abiteboul, S. (1997). Querying semi-structured data. In International Conference on Database Theory (pp. 1-18). Springer, Berlin, Heidelberg.

Agarwal, S., & Sureka, A. (2014, September). A focused crawler for mining hate and extremism promoting videos on YouTube. In Proceedings of the 25th ACM conference on Hypertext and social media (pp. 294-296).

Ahmadi-Abkenari, F., & Selamat, A. (2012). An architecture for a focused trend parallel Web crawler with the application of clickstream analysis. Information Sciences, 184(1), 266-281.

Ahmadi-Abkenari, F., & Selamat, A. (2012). An architecture for a focused trend parallel Web crawler with the application of clickstream analysis. Information Sciences, 184(1), 266-281.

Badawi, M., Mohamed, A., Hussein, A., & Gheith, M. (2013). Maintaining the search engine freshness using mobile agent. Egyptian Informatics Journal, 14(1), 27-36.

Badawi, M., Mohamed, A., Hussein, A., & Gheith, M. (2013). Maintaining the search engine freshness using mobile agent. Egyptian Informatics Journal, 14(1), 27-36.

Baykan, E., Henzinger, M., & Weber, I. (2013). A comprehensive study of techniques for URL-based web page language classification. ACM Transactions on the Web (TWEB), 7(1), 1-37.

Bošnjak, M., Oliveira, E., Martins, J., Mendes-Rodrigues, E., & Sarmento, L. (2012). Twitterecho: a distributed focused crawler to support open research with twitter data. In Proceedings of the 21st International Conference on World Wide Web (pp. 1233-1240).

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56(18), 3825-3833.

Chakrabarti, S., Van den Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. Computer networks, 31(11-16), 1623-1640.

Chen, Z., Liu, J., Zhai, H., Jiang, L., & Cao, B. (2012). Web Page Recognition Algorithm Based on Link Analysis in Theme Search Engine. In 2012 Second International Conference on Cloud and Green Computing (pp. 405-409). IEEE.

Chy, A. N., Seddiqui, M. H., & Das, S. (2014). Bangla news classification using naive Bayes classifier. In 16th Int'l Conf. Computer and Information Technology (pp. 366-371). IEEE.

Dong, H., & Hussain, F. K. (2012). Self-adaptive semantic focused crawler for mining services information discovery. IEEE Transactions on Industrial Informatics, 10(2), 1616-1626.

Dong, H., & Hussain, F. K. (2013). SOF: a semi-supervised ontology-learning-based focused crawler. Concurrency and Computation: Practice and Experience, 25(12), 1755-1770.

Du, Y., Hai, Y., Xie, C., & Wang, X. (2014). An approach for selecting seed URLs of focused crawler based on user-interest ontology. Applied Soft Computing, 14, 663-676.

Furche, T., Gottlob, G., Grasso, G., Guo, X., Orsi, G., & Schallhart, C. (2013). The ontological key: automatically understanding and integrating forms to access the deep Web. The VLDB Journal, 22(5), 615-640.

Gravano, L., Ipeirotis, P. G., & Sahami, M. (2003). QProber: A system for automatic classification of hidden-web databases. ACM Transactions on Information Systems (TOIS), 21(1), 1-41.

Hammer, J., & Fiedler, J. (2000). Using mobile crawlers to search the web efficiently. International Journal of Computer and Information Science, 1(1), 36-58.

Koster, M. (1994). A standard for robot exclusion. NEXOR.

Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of Web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1218.

Madaan, R., Dixit, A., Sharma, A. K., & Bhatia, K. K. (2010). A framework for incremental domain-specific hidden web crawler. In International Conference on Contemporary Computing (pp. 412-422). Springer, Berlin, Heidelberg.

Mali, S., & Meshram, B. B. (2011). Focused web crawler with revisit policy. In Proceedings of the International Conference & Workshop on Emerging Trends in Technology (pp. 474-479).

Nguyen, T. H., Nguyen, H., & Freire, J. (2010). PruSM: a prudent schema matching approach for web forms. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1385-1388).

Ning, H., Wu, H., He, Z., & Tan, Y. (2011). Focused crawler URL analysis model based on improved genetic algorithm. In 2011 IEEE International Conference on Mechatronics and Automation (pp. 2159-2164). IEEE.

Olston, C., & Najork, M. (2010). Web crawling. Now Publishers Inc.

Rattan, D., Bhatia, R., & Singh, M. (2013). Software clone detection: A systematic review. Information and Software Technology, 55(7), 1165-1199.

Singh, L., & Sharma, D. K. (2013). An approach for accessing data from hidden web using intelligent agent technology. In 2013 3rd IEEE International Advance Computing Conference (IACC) (pp. 800-805). IEEE.

Turek, W., Opalinski, A., & Kisiel-Dorohinicki, M. (2011). Extensible web crawler–towards multimedia material analysis. In International Conference on Multimedia Communications, Services and Security (pp. 183-190). Springer, Berlin, Heidelberg.

Uzun, E., Serdar Güner, E., Kılıçaslan, Y., Yerlikaya, T., & Agun, H. V. (2014). An effective and efficient Web content extractor for optimizing the crawling process. Software: Practice and Experience, 44(10), 1181-1199.

Yani Achsan, H. T., & Wibowo, W. C. (2013). A Fast Distributed Focused-Web Crawling. Annals of DAAAM & Proceedings, 24(1).

Yu, H. L., Bingwu, L., & Fang, Y. (2010). Similarity computation of web pages of focused crawler. In 2010 International Forum on Information Technology and Applications (Vol. 2, pp. 70-72). IEEE.

Zhang, Z., Dong, G., Peng, Z., & Yan, Z. (2011). A framework for incremental deep web crawler based on URL classification. In International Conference on Web Information Systems and Mining (pp. 302-310). Springer, Berlin, Heidelberg.

Zheng, Q., Wu, Z., Cheng, X., Jiang, L., & Liu, J. (2013). Learning to crawl deep web. Information Systems, 38(6), 801-819.