A Study of Focused Web Crawling Techniques

Gourav Shrivastava
Department of CSE, MANIT
Bhopal M. P., India
gashr83@gmail.com

Praveen Kaushik
Department of CSE, MANIT
Bhopal M. P., India
pk_kaushik@rediffmail.com

R. K. Pateriya
Department of CSE, MANIT
Bhopal M. P., India
pateriyark@gmail.com

***Abstract*—**In the recent years, the growth of data on the web is increasing exponentially. Due to this exponential growth, it is very crucial to find the accurate and significant information on the Web. Web crawlers are the tools or programs which find the web pages from the World Wide Web by following hyperlinks. Search engines indexes web pages which can be further retrieved by entering a query given by a user. The immense size and an assortment of the Web make it troublesome for any crawler to recover every pertinent information from the Web. In this way, different variations of Web crawling techniques are emerging as an active research area. In this paper, we survey the learnable focused crawlers.

Keywords- *Learnable Focused Crawler, Search Engine, Web Crawling.*

*****

I. INTRODUCTION

The search engines plays very vital role in every ones life today. In the large world wide Web, finding significant information is very critical task. Search engines utilize ranking algorithms which only gives the pages that matched as more relevant in based on the user query. A large number of users only visit the first few results which are relevant to the query they given. It is very important to order the results of user interest efficiently. A web crawler is a tool which explores the data on the Web. It continues visiting web pages on the internet to collect data that can be listed by an indexer to deal with any users query effectively. In general, there are two types of crawler's generic web crawlers and focused Web crawlers [1]. The generic web crawlers are not constrained to website pages of a specific subject or area. They continue following links unendingly and get all website pages they experience. While focused crawlers only visit the pages that are more relevant according to user query they do not follow all the links endlessly so the time taken by the focused crawlers are less as compared to the generic crawlers also the resources can be better utilized. First focused has been proposed by Chakrabarti et al [1] which only looks at web pages that are related to the predefined topics. Most of the focused crawling approaches are based on the Best-first-search (BFS) algorithm in which the frontier queue of a crawler is served as a priority queue and every URL in this queue has a ranking score based on ranking algorithms. The main aim of this paper is to study and review the different techniques used in the field of learnable focused web crawling.

Architecture of Web Crawler

The architecture of the basic Web crawler is shown in Figure 1.It start by taking seed URLs as an input and gives crawled WebPages as output. It starts by from a set of seed URLs, and then extract the links and downloads the webpage corresponds to URL, and then adds URLs or hyperlinks presents on webpage to create URL list and stores them in a queue to visit further. URL is chosen one by one from the list of URLs called frontier queue and then fetch a page corresponding to URL taken from the frontier. Crawler continues visiting URLs until frontier queue becomes empty.

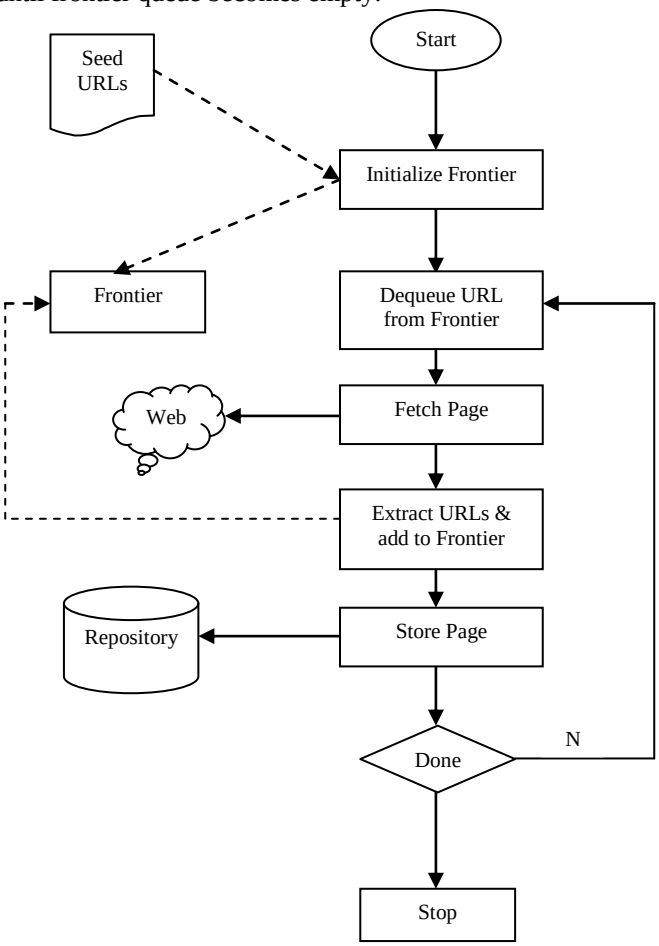

Figure 1 Basic Web Crawler Architecture

II. TYPES OF WEB CRAWLER

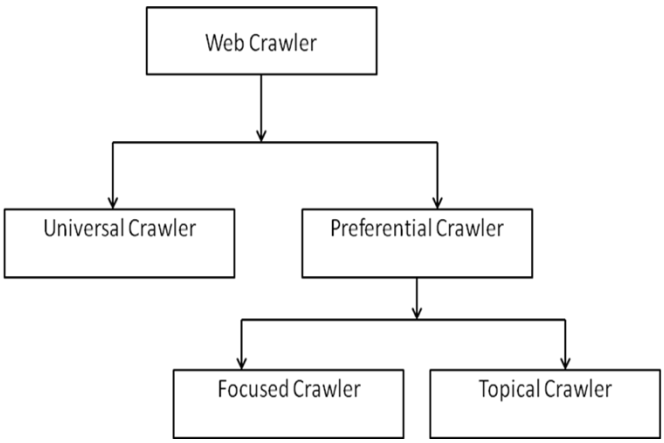

Figure 2- Classification of Web crawlers

Universal Crawler

These are general purpose Web crawlers and are not capable to crawl pages on a specific topic or domain. They keep visiting links continuously and download each web page by following seed URLs. Such types of crawlers are not effective and consume network bandwidth unnecessarily.

Preferential crawler

Preferential crawlers do not crawl each links they only crawl the links based on the user supplied criteria this can be a filter or a topic of interest. The user supplied criteria helps in guiding the crawler. Preferential crawlers are further categorized as focused and topical crawler. The first focused crawler proposed by Chakrabarti et al.[1] use the classifier to build a text classifier using some labelled example set. Then the classifier would guide the crawler by preferentially selecting from frontier those pages that appear most likely to belong to categories of interest, according to classifier's prediction. In topical crawler or topic-specific crawler pages are visited at query time only they do not have text classifiers to guiding the crawl. Topical crawling is suitable for applications that look for very recently posted documents, which a search engine may not have indexed yet. Topical crawling crawls the pages based on the specific topic or criteria while in focused crawling is based on the classification by some labeled examples of relevant and non-relevant web pages.

Focused Web Crawler

Focused crawlers only looks for the pages that are relevant to a predefined topic and download the same based on some criteria to efficiently crawl the web to minimize the wastage of resources like time, storage and network bandwidth. Focused crawlers fulfill the topic specific needs of a user or an organization to maintain a good quality up-to-date database desired by them. Focused crawlers only downloads the pages of interest minimize the downloading of not relevant pages[1]. Classification techniques are used to determine the order of a web page on a particular topic. Basically focused crawlers' uses two techniques, page scoring and link scoring. Page scoring determines whether a web page is relevant to a given topic or not. Link scoring determines the links to be crawled, and prioritize the order of the links to be crawled. Classification algorithms are generally used in page scoring. They first download a page and then use the classifier to decide whether the page is about the topic or not[2]. Focused crawling starts with seed URLs as input, fetch and extract pages by following hyperlinks but restricted to download the webpage which meets some predefined matching or relevancy criteria and store in a local repository. Crawler continue visiting webpages until a sufficient number is achieved or it meets a stopping criteria.

III. CHALLENGES IN FOCUSED WEB CRAWLING

The researchers face some challenges while working on Web crawlers. Some challenges related to crawling the web are as follows.

Missing relevant pages-

The main issue with focused crawling in which the crawler miss some relevant pages.

Non-uniform structures of the web-

The Web structure is dynamic in nature which is due to the inconsistency of data structures used. There is no uniform norms are available for building a webpage. The absence of uniformity can affect the collection of data because the data in the web is semi-structured or unstructured.

Freshness-

The objective of a Web crawler must be to download the updated pages to maintain the freshness of the database.

Multimedia Content Crawling -

A web crawler can conveniently analyze text but not the multimedia. Analyzing multimedia is an open issue.

Deep Web Crawling-

Crawling the deep web is another challenge because it remains unavailable for standard Web crawlers. The deep web or hidden webs are parts of the World Wide Web whose contents are not indexed by standard web search engines.

Network Bandwidth and Impact on Web Servers-

Crawling a large number of pages can significantly reduce the network bandwidth and increases the load on web servers. It is desirable to retrieve only highly related pages so that the underlying resources are not overloaded.

IV. LEARNABLE FOCUSED CRAWLER

The main function of focused crawler is to find relevant webpages by following seed URLs based on any particular topic given by the user. Learnable focused crawler can be trained through supervised, unsupervised or some other learning techniques to further guide the crawler to find more and more relevant pages on a particular topic.For The main function of focused crawler is to find relevant webpages by following seed URLs based on any particular topic given by the user. Learnable focused crawler can be trained through supervised, unsupervised or some other learning techniques to further guide the crawler to find more and more relevant pages on a particular topic.For learning a crawler can use various datasets like 20 News group, Reuters or online library like Open Directory Project (Dmoz). Techniques like SVM, Naive Bayes etc are used for classification. Table 1 gives a comparative study of various learnable focused crawlers which provides brief overview of their strengths and limitations which enable researchers to find further direction to work on this.

Table 1- Comparative study of learnable focused crawlers

Article	Technique used	Strength	Limitation	Implementation Language
Huang and Ye 2004[3]	on-line incremental learning using SVM and Naive Bayes	• Can starts with a few sample web pages	• As a performance metric only harvest rate is used	Java
A. Rungsawang et al. 2005 [4]	Consecutive crawling learned from previous crawls	• Collects more relevant pages using past learning experience	• Works with ODP directory only. • Less number of web pages examined	C
A. Ghozia et al. 2008[5]	Bayesian Object Based Approach using supervised learning	• Crawler focused to user interest towards the topic	• Works better for the if website has a well defined contents	Java
Zhang and Lu 2010 [6]	semi-supervised clustering approach and Fuzzy logic	• Online learning capability	• Not performed well in initial stage of crawling	--
Mejdl S. Safran et al. 2012[7]	Naive Bayes Learning	• Improved accuracy of relevancy prediction	• Use only four relevance attributes to predict the relevance of unvisited URLs.	Java
Kumar and Vig 2012[8]	Hub score is used as a learning parameter for the crawler to select best seed pages	• improvement in the precision value for the crawler	• Performs better after consecutive crawls	--
Rong qian et al. 2013[9]	Based on neural network &reinforcement learning	• improves the efficiency and accuracy	• Crawling speed slows down significantly when taking too many parameters in to account	--
Zheng et al 2013.[10]	Deepweb crawling based on the reinforcement learning (RL)	• Suggests applicability of different crawling methods for different deep websites	• Not considered how to apply the RL method to crawl deep web databases by querying multiple attributes	--
Peng and Liu 2013[11]	Heuristic based approach based on CBP-SLC and SVM for classification	• Identifies more reliable negative documents from the unlabeled example set	• Tidy up the web pages is required	Java

V. CONCLUSION

A focused crawler only search and downloads the web pages which are relevant to the search topic given to the crawler. A learning-based focused crawler has learning ability to adapt to its search topic and to improve the accuracy of its prediction of relevancy of unvisited URLs. Learning based crawlers uses classification algorithms to guide the crawlers. In this paper, we review the various focused learnable crawler approaches.

REFERENCES

[1] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: a new approach to top-specific Web source discovery," *Comput. Networks*, vol. 31, no. 11–16, pp. 1623–1640, 1999.

[2] D. Taylan, M. Poyraz, S. Akyokuş, and M. C. Ganiz, "Intelligent focused crawler: Learning which links to crawl," *INISTA 2011 - 2011 Int. Symp. Innov. Intell. Syst. Appl.*, pp. 504–508, 2011.

[3] D. Library, "wHunter : A Focused Web Crawler – A Tool for," no. 60221120145, pp. 519–522, 2004.

[4] A. Rungsawang and N. Angkawattanawit, "Learnable topic-specific web crawler," vol. 28, pp. 97–114, 2005.

[5] A. Ghozia, H. Sorour, and A. Aboshosha, "25 NATIONAL RADIO SCIENCE CONFERENCE (NRSC 2008) IMPROVED FOCUSED CRAWLING USING BAYESIAN OBJECT BASED APPROACH Ahmed Ghozia and Hoda Sorour Faculty of Electronic Eng ., Menofiya University The rapid growth of the World-Wide-Web made it difficult for ge," no. Nrsc, 2008.

[6] H. Zhang and J. Lu, "SCTWC : An online semi-supervised clustering approach to topical web crawlers," vol. 10, pp. 490–495, 2010.

[7] M. S. Safran, A. Althagafi, and D. Che, "Improving Relevance Prediction for Focused Web Crawlers," 2012.

[8] M. Kumar and R. Vig, "Learnable Focused Meta Crawling Through Web," *Procedia Technol.*, vol. 6, no. 1994, pp. 606–611, 2012.

[9] G. Zhao, "A Topic-specific Web Crawler Based on Content and Structure Mining," pp. 458–461, 2013.

[10] Q. Zheng, Z. Wu, X. Cheng, L. Jiang, and J. Liu, "Learning to crawl deep web," *Inf. Syst.*, vol. 38, no. 6, pp. 801–819, 2013.

[11] T. Peng and L. Liu, "Knowledge-Based Systems Focused crawling enhanced by CBP – SLC," vol. 51, pp. 15–26, 2013.