International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

Focused Web Crawler and its Approaches

Jay Sampat[Á], Anmol Jain[Á*] and Dharmeshkumar Mistry[Á]

[Á]Department of Computer Science, Dwarkadas J Sanghvi College of Engineering, Vile Parle(W), Mumbai, India

Accepted 05 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

There has been a rapid growth of the world-wide web which has scaled beyond our imaginations. To surmount these challenges search engines are used. One of the most important type of crawler is Focused crawler which is used to index information according to a particular topic. To maximize the possibility of downloading relevant documents focused crawler makes a prediction of hyperlinks visiting priority which in turn helps to reduce downloading of irrelevant documents and drastically saves network resources and hardware. Instead of using keywords topics are specified by using commendable documents. One of the most important feature of this type of web crawler is collecting and indexing all accessible web credentials. This crawler mainly diagnosis its crawl boundary to search different URLs. In this paper we'll illustrate a clear cut comparison between focused and standard web crawlers as well as various approaches of focused crawling like contextual and priority based crawling.

Keywords*: web crawlers, focused crawlers, web pages, priority based, contextual based, indexing.*

1. Introduction

There are instances where a search engine need not be completely general purpose and cover every possible site on the Web. Sometimes it can be focused on particular topics that may interest only a specific set of users. Let's consider an example of a crawler that provides information to a specialized portal like sports, hence it could be programmed to ignore sites with content in different topics such as health and movies. A focused crawler can be defined as a topic sensitive crawler that looks for data related to certain topics only. The number of sites in a particular domain can be small enough so that a focused crawler can download them in a relatively short span of time.

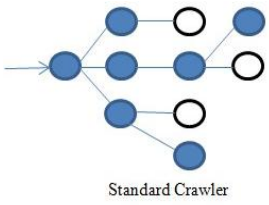

Fig 1

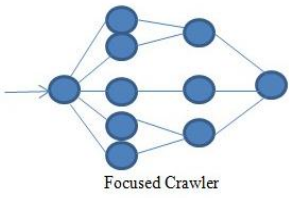

Fig. 2

A focused crawler is web crawler that efficiently gathers Web pages that fulfills a specific criteria, by carefully prioritizing the crawl frontiers. Crawl frontier is the link on a web page that a web crawler can select while performing crawling process. Some predicates may be based on simple or surface properties. A selective crawler's task may be to crawl pages from only the .in domain whereas some other crawler's aim would be to crawl pages only from .jp domain. While other predicates may be a bit comparative, e.g., crawl pages with large Page Rank or Page Hit, or crawl pages that are related to cricket.

2. Comparison

Table 1 Difference between a standard web crawler and focused web crawler

Crawler	General Web Crawler	Focused Web Crawler
Found by	Different sources and contributors	Chakrabarti, Berg and Dom
Definition	It locates information on WWW, indexes all the words in a document and follows each and every hyperlink.	It indexes information based on Domains, Applications and the query that's inserted.
Web pages	May or may not be related to each other	Has to be related to a certain domain.
Relevance	Less relevant web pages are discovered.	More relevant web pages are discovered.
Consumption of resources	Less	High
Performance dependency	Independent	Dependent on link richness within a specific domain
Performance	Low	High

3. Different approaches of Focused Web Crawler

Focused crawler approaches are usually categorized with respect to their dependency on determining the relevant web pages they focus on. They are:

*Corresponding author: **Anmol Jain**

- Priority based focused crawler
- Contextual based focused crawler
- Structure based focused crawler

3.1 Priority based Focused Crawler

This type of crawler keeps the relative score along with the corresponding URL's to be visited in a priority queue. When a URL from a priority queue is deleted, it returns a maximum score URL. This type shows about 87% improved results over simple crawler. One drawback of this method is that it is very time consuming and this con can be eliminated by implementing parallel algorithm.

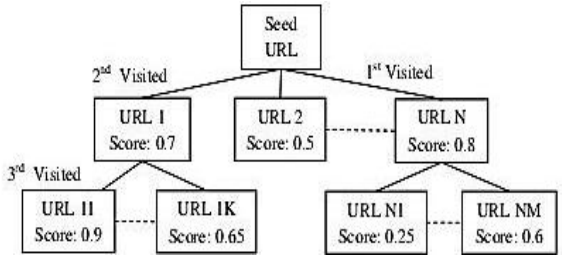

Fig. 3

3.2 Contextual based focused crawler

It is a different approach that address the very problem of assigning the right importance to various web pages along a crawl path. If there is a set of documents by querying some search engines we will be able to build a representation of web pages that occur within a particular link distance of those documents. If the Web is backcrawled; these analyzed pages will point to the set of interesting and relevant documents or web pages. All the information that is retrieved is stored in a structure named context graph. These context graphs maintains the related distance for each page. This related distance is nothing but minimum number of necessary links so that one can traverse back to reach the initial set.

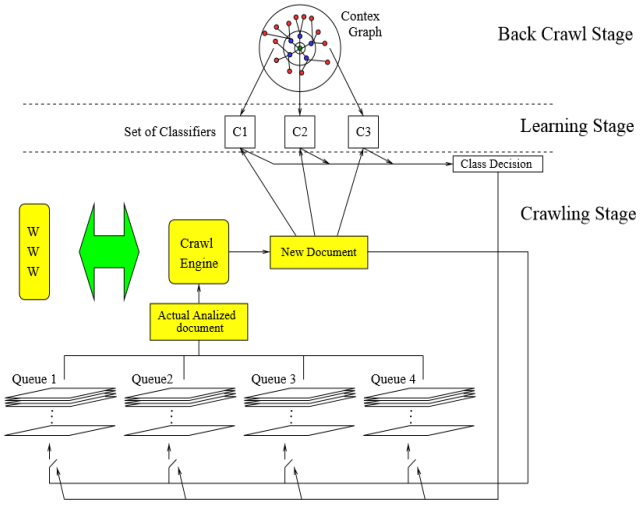

Fig. 4

4. Workflow of focus based crawler

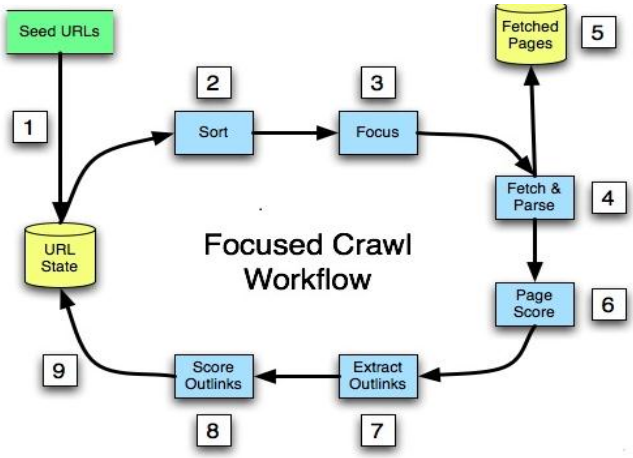

Fig. 5

The workflow of focused crawler is discussed below:

1. Firstly URL state database is loaded along with an initial set of URL's.
2. After this the first loop of iteration in the focussed crawl can begin. This iteration extracts the unprocessed URLs and will sort them in an order according to their link score.
3. The next step is very critical. A decision has to be made about the number of URLs to be processed in the loop.
4. As soon as the set of accepted URLs has been created, the standard fetch process initiates. Therefore, all the usual steps required for efficient fetching are done.
5. These fetched pages are also saved into the Fetched Pages database.
6. The next critical step is to give the parsed page content a page score which is nothing but the value representing the proximity of the page match. Usually this is a value from 0.0 to 1.0, with higher scores being better.
7. After the scoring is done, every outlink which is found in the parse is extracted.
8. Finally we divide the score for the page among all the outlinks.
9. Then we update the URL State database with the results of fetch attempts i.e. succeeded or failed. After that we add all the newly discovered URLs.

Conclusion

Web crawling is one of the main component in applications like search engines. We compared between standard and focused web crawlers to understand which one is better and also discussed about the merits of various approaches like priority based as well as contextual based focused crawling. The advantages of focused crawler are that we spend less money, time & effort processing web pages that are most unlikely to be of value or worth. Also the manner in which focused crawler proceeds has been illustrated and documented in this paper.

Acknowledgement

We would like to thank our honourable principal Dr. Hari Vasudevan of D. J. Sanghvi College of Engineering and

Head of Department of Computer Engineering, Dr. Narendra Shekhokar for giving us the facilities and providing us with a propitious environment for working in college.Also, we would like to thank Prof. Dharmeshkumar Mistry, without whose guidance and expertise this paper wouldn't have been possible.

References

Baeza-Yates, Ricardo. Applications of Web Query Mining. Springer, 2005: 7-22.

Batsakis, Sotiris, Euripides Petrakis, and Evangelos Milios. Improving the performance of focused web crawlers. Elsevier, 2009.

Castillo, Carlos. EffectiveWeb Crawling. ACM, 2005 .

Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. Elsevier, 1999.

Diligenti, Coetzee, Lawrence, Giles, and Gori. Focused Crawling Using Context Graphs. 26th International Conference on Very Large Databases, VLDB 2000. Cairo, Egypt, 2000. 527–534.

Karkaletsis, Vangelis, Konstantinos Stamatakis, James Horlock, Claire Grover, and James R. Curran. DomainSpecificWeb Site Identification: The CROSSMARC Focused Web Crawler. Proceedings of the 2nd International Workshop on Web Document Analysis (WDA2003). Edinburgh, UK, 2003.

Liu, Hongyu, and Evangelos Milios. Probabilistic Models for Focused Web Crawling. Computational Intelligence, 2010.

Liu, Hongyu, Evangelos Milios, and Larry Korba. Exploiting Multiple Features with MEMMs for Focused Web Crawling. NRC, 2008.

Rungsawang, Arnon, and Niran Angkawattanawit. Learnable topic-specific web crawler. Science Direct, 2005: 97–114.

Suel, Torsten, and Vladislav Shkapenyuk. Design and Implementation of a High-Performance Distributed Web Crawler. Proceedings of the IEEE International Conference on Data Engineering. 2002.