

DESIGNING A CRAWLER AS AN INTELLIGENT AGENT

Jahman N'Daw

Class of 2007

#C2495

570-577-6577

jndaw@bucknell.edu

123-68-7806

Faculty Advisor: Professor Xiannong Meng, Computer Science

X_____

Jahman N'Daw

X_____

Xiannong Meng

PURPOSE OF THE PROJECT:

The purpose of this project is to design and implement a crawler that will process relevant information on the web. After its integration to the search engine, the crawler will provide web users a powerful educational search tool. It will also enable users to access resourceful educational intranets such as universities and libraries. On the other hand, it will empower these intranets to establish communications, as well as opening opportunities for future collaborations among them.

PROJECT DETAILS:

Presently, the biggest advantage for the users of the Internet is the immense amount of information available on the World Wide Web (www). However, the biggest drawback is the difficulty of finding relevant information. Search engines provide only a partial solution to this problem. The enormous amount of information available on the Web perpetually limits the efficiency, speed and accuracy of search engines. This limitation is becoming problematic since the information available on the Web is growing constantly and rapidly at an accelerated rate[1]. There are billions of web pages with their number increasing at a rate of a million pages per day[2] and 40% of all web pages changing weekly[3, 4]. Google, for instance, provides a search of over 3 billion web pages[5] which represents only a part of whole the web. Therefore, it is important to address the problems of limited efficiency, speed and accuracy. This project is an attempt to improve information retrieval on the web.

Crawlers are intelligent agents. They are also known as spiders or bots. They are specialized programs that automatically visit sites and index the web pages by creating entries in the databases of search engines. They do so by "crawling" through a site a page at a time, following the links to other pages on the site until all pages have been read[4]. However, maintaining currentness of indices by constant crawling is rapidly becoming impossible due to the increasing size and dynamic content of the web[6]. The main reason is that current crawlers can easily wander off the targeted web sites when they follow hyperlinks.

The challenge of this project is to design and implement a crawler that will stay focused when crawling. A domain-focused crawler, that is integrated to a search engine, selectively seeks out pages that are relevant to pre-defined topics[2,3]. Designed as such, the crawler will no longer target the entire web. It will only crawl sets of hosts that are in a given domain (e.g. *.edu). It will crawl particular topical sections of the web without exploring the irrelevant ones[7]. This behavior is called "an intelligent crawling"[8]. A crawler that is designed to produce efficient, speedy, and accurate information is a major educational tool for researchers, academicians, and students. Such users will have access to the potentially rich intranets and will be able to locate wide range of information. In short, the applicability of the proposed crawler will allow communications and collaborations between these intranets.

Crawlers are intelligent multi-purpose tools and their range of parameters can be reconfigured. Certain parameters pertain to the speed at which the crawler downloads pages. The walking and crawling characteristic of the crawler, the design that allows it to only go after indexed pages, the schedule of its activity, the domains that it connect to and the pages that it will accept from a single domain as well as the frequency of its activity are other attributing parameters of a crawler. The design and implementation of a crawler which will optimize all these parameters, and improve the speed and efficiency of the search engine and the accuracy of the search results. The outcome of the proposed project will be a powerful research tool, that is a domain-focused crawler which will provide a best user experience.

METHODOLOGY:

This research will take place in Bucknell University during June and July of 2004. The time table of the project will have two sequences.

During the first two weeks of June, I will complete a research on relevant information. I will start researching the material after the mid-term examinations of the Spring semester. I will work closely with Dr. Xiannong Meng everyday of the week and examine how different implementations of crawlers work for a small set of documents. The design of the crawler will be selected after analyzing various possibilities. During the last two weeks of June, Dr. Meng and I will work on specifying all aspects of the selected design. A precise documentation of the stages of this process will be provided with precision and details. During this period, we will also consider relevant theoretical approaches to analyze and minimize possible risks, as well as consider best methodologies to identify and manage problems, in order to provide a correct algorithm to implement the crawler[9].

During July, we will be implementing the crawler by transforming the algorithm into a particular programming language. After completing the implementation processes, we will test the code, remove all logical errors and design a set of data to test the crawler's performance. We will refine the crawler in stages: first, by developing a simplified working program for testing purposes; then, refining the solution adding new features. The outcome of this strategy will be a well-structured crawler that is integrated into the search engine and is a useful educational research tool. Our crawler will enable the search engine to produce accurate information. Resulting efficient, speedy and accurate access to resourceful intranets will establish communications and collaborations between numerous intranets creating a "network of networks". Throughout this research, I will produce summaries of the relevant journal articles, conference papers, white papers and books. I will discuss my findings with Dr. Meng on regular basis. During the period of designing the proposed crawler, I will test different algorithms and measure their performances.

Not only will this multi-sided Summer Research at Bucknell allow me to gain research experiences, but also is critical for expanding my knowledge on design, deployment, and performance of web crawlers. All these activities will provide me with essential experiences and needed knowledge to become an accomplished computer engineer. I eagerly anticipate having the opportunity to become a research assistant for this project.

FACULTY ENDORSEMENT:

I am very excited about the prospective that I may have an opportunity to work with Jahman N'Daw in a summer research project about focused web crawler. I have been working in the area of intelligent web search for a few years. My colleagues and I have published a number of journal papers, book chapters and conference proceeding papers in the area. However we often come to a roadblock that we don't have our own web crawler to collect information. We had to rely on the information collected by other search engines in our empirical study. We have been dreaming of having a crawler of our own that can integrate with the search engines we use in our experiments.

Jahman comes at a right time with the right project! I had Jahman in my Introduction to Computer Science II course. Though only a sophomore, he impressed me the most with his enthusiasm for research in the area of Internet and World Wide Web, his eagerness to explore the unknowns, and his ability to keep himself focused on the project.

This crawler project will give us the opportunity to experiment with the ideas we have been thinking about that we couldn't implement because of the lack of time. We would like to focus crawling by following only the web servers in a specific domain. When the crawler is built as proposed, it will allow us to further investigate various web search issues such as personalized search, clustering the search results, and further, federated web search. Jahman will certainly benefit from the project in various aspects. He will learn first hand how research is conducted, how to do literature search, how to design and implement a relatively large software, how to test and verify hypotheses after the implementation is done, and how to write scientific research papers. I very much look forward to working with Jahman in the coming summer!

REFERENCES:

- [1] Jenny Edwards, Kevin McCurley, John Tomlin, *"An Adaptive Model for Optimizing Performance of an Incremental Web Crawler"* Proceedings of the 10th International World Wide Web Conference, Hong Kong, May 2001.
- [2] Chakrabarti, S., van der Berg, M., & Dom, B. *"Focused crawling: a new approach to topic-specific Web resource discovery."* Proceedings of 8th International World Wide Web Conference, Toronto, Canada, May, 1999.
- [3] Dennis Fetterly, Mark Manasse, Marc Najork, Janet L. Wiener *"A large-scale study of the evolution of Web pages."* Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, May 2003.
- [4] Andrei Z. Broder, Marc Najork, Janet L. Wiener, *"Efficient URL Caching for World Wide Web Crawling."* Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, May 2003.
- [5] The Google Search Engine, <http://www.google.com>

[6] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "*Focused crawling using context graphs.*" Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May 2000.

[7] Soumen Chakrabarti, Kunal Punera, Mallela Subramanyam, "*Accelerated Focused Crawling through Online Relevance Feedback* " Proceedings of the 11th International World Wide Web Conference, Honolulu, Hawaii, May 2002.

[8] Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. "*Intelligent crawling on the World Wide Web with arbitrary predicates.*" Proceedings of the 10th International World Wide Web Conference, Hong Kong, May 2001.

[9] Frank M. Carrano, Janet J. Prichard, "*Data Abstraction and Problem Solving with C++, Walls and Mirrors*", New York, Addison Wesley, 2002.