Effective Web Crawling

by

Carlos Castillo

Submitted to the University of Chile in fulfillment
of the thesis requirement to obtain the degree of
Ph.D. in Computer Science

Advisor	**Dr. Ricardo Baeza-Yates** University of Chile
Committee	Dr. Mauricio Marin University of Magallanes, Chile
	Dr. Alistair Moffat University of Melbourne, Australia
	Dr. Gonzalo Navarro University of Chile
	Dr. Nivio Ziviani Federal University of Minas Gerais, Brazil

This work has been partially funded by the Millennium Nucleus "Center for Web Research"
of the Millennium Program, Ministry of Planning and Cooperation – Government of Chile

Dept. of Computer Science - University of Chile

November 2004

Abstract

The key factors for the success of the World Wide Web are its large size and the lack of a centralized control over its contents. Both issues are also the most important source of problems for locating information. The Web is a context in which traditional Information Retrieval methods are challenged, and given the volume of the Web and its speed of change, the coverage of modern search engines is relatively small. Moreover, the distribution of quality is very skewed, and interesting pages are scarce in comparison with the rest of the content.

Web crawling is the process used by search engines to collect pages from the Web. This thesis studies Web crawling at several different levels, ranging from the long-term goal of crawling important pages first, to the short-term goal of using the network connectivity efficiently, including implementation issues that are essential for crawling in practice.

We start by designing a new model and architecture for a Web crawler that tightly integrates the crawler with the rest of the search engine, providing access to the metadata and links of the documents that can be used to guide the crawling process effectively. We implement this design in the WIRE project as an efficient Web crawler that provides an experimental framework for this research. In fact, we have used our crawler to characterize the Chilean Web, using the results as feedback to improve the crawler design.

We argue that the number of pages on the Web can be considered infinite, and given that a Web crawler cannot download all the pages, it is important to capture the most important ones as early as possible during the crawling process. We propose, study, and implement algorithms for achieving this goal, showing that we can crawl 50% of a large Web collection and capture 80% of its total Pagerank value in both simulated and real Web environments.

We also model and study user browsing behavior in Web sites, concluding that it is not necessary to go deeper than five levels from the home page to capture most of the pages actually visited by people, and support this conclusion with log analysis of several Web sites. We also propose several mechanisms for server cooperation to reduce network traffic and improve the representation of a Web page in a search engine with the help of Web site managers.

Publications related to this thesis

The crawling model and architecture described in Chapter **??** was presented in the second Hybrid Intelligent Systems conference [BYC02] (HIS 2002, proceedings published by IOS Press), and introduced before in preliminary form in the eleventh World Wide Web conference [CBY02].

The analysis and comparison of scheduling algorithms, in terms of long-term and short-term scheduling in Chapter **??** was presented in the second Latin American Web conference [CMRBY04] (LA-WEB 2004, published by IEEE CS Press).

The model and analysis of browsing behavior on the "Infinite Web" on Chapter **??** was presented in the third Workshop on Algorithms and Models for the Web-Graph [**?**] (WAW 2004, published by Springer LNCS).

Most of the proposals about Web server cooperation shown in Chapter **??** were introduced in preliminary form in the first Latin American Web conference [**?**] (LA-WEB 2003, published by IEEE CS Press).

Portions of the studies on Web structure and dynamics shown in Chapter **??** appear as a chapter in the book "Web Dynamics" [BYCSJ04] (published by Springer), and were presented in the 8th and 9th String Processing and Information Retrieval conferences [BYC01, BYSJC02] (SPIRE 2001, published by IEEE CS Press and SPIRE 2002, published by Springer LNCS).

An application of the WIRE crawler to characterize images, not described in this thesis, was presented in the first Latin American Web conference [JdSV$^+$03] (LA-WEB 2003, published by IEEE CS Press) and the third Conference on Image and Video Retrieval [BYdSV$^+$04] (CIVR 2004, published by Springer LNCS).

The WIRE crawler developed during this thesis is available under the GNU public license, and can be freely downloaded at `http://www.cwr.cl/projects/WIRE/`. The user manual, including step-by-step instructions on how to use the crawler, is available at the same address.

Acknowledgements

This thesis would not have been possible without **O.P.M.** During the thesis I received mostly the financial support of grant P01-029F of the Millennium Scientific Initiative, Mideplan, Chile. I also received financial support from the Faculty of Engineering and the Computer Science Department of the University of Chile, among other sources.

What you are is a consequence of whom you interact with, but just saying "thanks everyone for everything" would be wasting this opportunity. I have been very lucky of interacting with really great people, even if some times I am prepared to understand just a small fraction of what they have to teach me. I am sincerely grateful for the support given by my advisor **Ricardo Baeza-Yates** during this thesis. The comments received from the committee members **Gonzalo Navarro**, Alistair Moffat, Nivio Ziviani and Mauricio Marin during the review process were also very helpful and detailed. For writing the thesis, I also received data, comments and advice from Efthimis Efthimiadis, Marina Buzzi, Patrizia Andrónico, Massimo Santini, Andrea Rodríguez and Luc Devroye. I also thank Susana Docmac and everybody at **Newtenberg**.

This thesis is just a step on a very long road. I want to thank the professors I met during graduate studies: **Vicente López**, Claudio Gutierrez and José Pino; also, I was lucky to have really inspiring professors during the undergraduate studies: Martin Matamala, Marcos Kiwi, Patricio Poblete, Patricio Felmer and **José Flores**. There were some teachers in high and grade school that trusted in me and helped me get the most out of what I was given. During high school: Domingo Almendras, Belfor Aguayo, and in grade school: Manuel Guiñez, Ivonne Saintard and specially **Carmen Tapia**.

I would said at the end that I owe everything to my parents, but that would imply that they also owe everything to their parents and so on, creating an infinite recursion that is outside the context of this work. Therefore, I thank **Myriam** and **Juan Carlos** for being with me even from before the beginning, and sometimes giving everything they have and more. I am also thankful for the support of all the members of my family, specially Mercedes Pincheira.

Finally, my beloved wife **Fabiola** was exactly 10,000 days old on the day I gave my dissertation, and I need no calculation to say that she has given me the best part of those days – thank you.

Contents

List of Figures

List of Tables

Bibliography

[BYC01] Ricardo Baeza-Yates and Carlos Castillo. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval*, pages 21–32, Laguna San Rafael, Chile, November 2001. IEEE Cs. Press.

[BYC02] Ricardo Baeza-Yates and Carlos Castillo. Balancing volume, quality and freshness in web crawling. In *Soft Computing Systems - Design, Management and Applications*, pages 565–572, Santiago, Chile, 2002. IOS Press Amsterdam.

[BYCSJ04] Ricardo Baeza-Yates, Carlos Castillo, and Felipe Saint-Jean. *Web Dynamics*, chapter Web Dynamics, Structure and Page Quality, pages 93–109. Springer, 2004.

[BYdSV+04] Ricardo A. Baeza-Yates, Javier Ruiz del Solar, Rodrigo Verschae, Carlos Castillo, and Carlos A. Hurtado. Content-based image retrieval and characterization on specific Web collections. In *Third international conference on image and video retrieval (CIVR)*, pages 189–198, Dublin, Ireland, July 2004. Springer LNCS.

[BYSJC02] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, dynamics and page quality. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 117 – 132, Lisbon, Portugal, 2002. Springer LNCS.

[CBY02] Carlos Castillo and Ricardo Baeza-Yates. A new crawling model. In *Poster proceedings of the eleventh conference on World Wide Web*, Honolulu, Hawaii, USA, May 2002. (Extended Poster).

[CMRBY04] Carlos Castillo, Mauricio Marin, Andrea Rodriguez, and Ricardo Baeza-Yates. Scheduling algorithms for Web crawling. In *Latin American Web Conference (WebMedia/LA-WEB)*, Riberao Preto, Brazil, 2004. IEEE Cs. Press. (To appear).

[JdSV+03] A. Jaimes, J. Ruiz del Solar, R. Verschae, D. Yaksic, R. Baeza-Yates, E. Davis, and C. Castillo. On the image content of the Chilean Web. In *Proceedings of Latin American Conference on World Wide Web (LA-WEB)*, pages 72–83, Santiago, Chile, 2003. IEEE Cs. Press.