DOCTORAL ABSTRACT

Effective Web Crawling

http://www.cwr.cl/projects/WIRE/

Carlos Castillo
Center for Web Research
University of Chile
ccastill@dcc.uchile.cl

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 [2, 8].

We implement this design in the WIRE project as an efficient Web crawler that provides an experimental framework for this research. The WIRE crawler developed during this thesis is available under the GNU public license, and can be freely downloaded from the URL specified above. The user manual, including step-by-step instructions on how to use the crawler, is available at the same address.

We have used our crawler to characterize the Chilean Web, using the results as feedback to improve the crawler design [1, 5, 4]. We have also used the crawler for content-based analysis of images [10, 6].

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 account for 80% of the total quality in both simulated and real Web environments [9].

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 [3].

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 [7].

Advisor: Ricardo Baeza-Yates (University of Chile). **Committee:** Mauricio Marín (University of Magallanes, Chile), Alistair Moffat (University of Melbourne), Gonzalo Navarro (University of Chile), Nivio Ziviani (Federal University of Minas Gerais, Brazil). **Dissertation date:** November 29, 2004.

References

[1] 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.

[2] 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.

[3] Ricardo Baeza-Yates and Carlos Castillo. Crawling the infinite Web: five levels are enough. In *Proceedings of the third Workshop on Web Graphs (WAW)*, volume 3243 of *Lecture Notes in Computer Science*, pages 156–167, Rome, Italy, October 2004. Springer.

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

[5] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, dynamics and page quality. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, volume 2476 of *Lecture Notes in Computer Science*, pages 117 – 132, Lisbon, Portugal, 2002. Springer.

[6] 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)*, volume 3115 of *Lecture Notes in Computer Science*, pages 189–198, Dublin, Ireland, July 2004. Springer.

[7] Carlos Castillo. Cooperation schemes between a web server and a web search engine. In *Proceedings of Latin American Conference on World Wide Web (LA-WEB)*, pages 212–213, Santiago, Chile, 2003. IEEE CS Press.

[8] 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).

[9] Carlos Castillo, Mauricio Marin, Andrea Rodríguez, and Ricardo Baeza-Yates. Scheduling algorithms for Web crawling. In *Latin American Web Conference (WebMedia/LA-WEB)*, pages 10–17, Riberao Preto, Brazil, October 2004. IEEE CS Press.

[10] A. Jaimes, J. Ruiz del Solar, R. Verschae, R. Baeza-Yates, C. Castillo, D. Yaksic, and E. Davis. On the image content of a Web segment: Chile as a case study. *Journal of Web Engineering*, 3(2):153–168, 2004.