Chapter 1

Introduction

This thesis is about Web crawling, the process used by Web search engines to download pages from the Web. This opening chapter starts with the main motivations for studying this process in Section 1.1. Section 1.3 introduces the WIRE project, the system that we use as a context for working on this topic. Section 1.4 explains the scope and organization of our work.

1.1 Motivation

1.1.1 From organic to mineral memory

As technology advances, concerns arise about how the new inventions may impair human capabilities. Plato, in his dialogue *Phaedrus*, tells the story of Theuth (Hermes) presenting his inventions to Pharaoh Thamus, who dislikes the idea of writing:

> "This, said Theuth, will make the Egyptians wiser and will give them better memories; it is a specific for both the memory and for the wit. Thamus replied: Oh most ingenious Theuth (...) this discovery of yours will create forgetfulness in the learners souls, because they will not use their memories; they will trust to the external written characters and not remember of themselves, (...) they will be hearers of many things and will have learned nothing; they will appear to be omniscient and will generally know nothing; they will be tiresome company, having the show of wisdom without the reality." [PlaBC]

Thamus considers that writing is a bad invention because it replaces human memory. There are others who consider writing as just an extension of the human memory, such as Umberto Eco, who in November 2003 gave a lecture about the future of books at the newly opened Library of Alexandria in Egypt:

> "We have three types of memory. The first one is organic, which is the memory made of flesh and

> blood and the one administrated by our brain. The second is mineral, and in this sense mankind has known two kinds of mineral memory: millennia ago, this was the memory represented by clay tablets and obelisks, pretty well known in this country, on which people carved their texts. However, this second type is also the electronic memory of today's computers, based upon silicon. We have also known another kind of memory, the vegetal one, the one represented by the first papyruses, again well known in this country, and then on books, made of paper." [Eco03]

The World Wide Web, a vast mineral memory, has become in a few years the largest cultural endeavour of all times, equivalent in importance to the first Library of Alexandria. How was the ancient library created? This is one version of the story:

> "By decree of Ptolemy III of Egypt, all visitors to the city were required to surrender all books and scrolls in their possession; these writings were then swiftly copied by official scribes. The originals were put into the Library, and the copies were delivered to the previous owners. While encroaching on the rights of the traveler or merchant, it also helped to create a reservoir of books in the relatively new city." [wik04]

The main difference between the Library of Alexandria and the Web is not that one was vegetal, made of scrolls and ink, and the other one is mineral, made of cables and digital signals. The main difference is that while in the Library books were copied by hand, most of the information on the Web has been reviewed only once, by its author, at the time of writing.

Also, modern mineral memory allows fast reproduction of the work, with no human effort. The cost of disseminating content is lower due to new technologies, and has been decreasing substantially from oral tradition to writing, and then from printing and the press to electronic communications. This has generated much more information than we can handle.

1.1.2 The problem of abundance

The signal-to-noise ratio of the products of human culture is remarkably high: mass media, including the press, radio and cable networks, provide strong evidence of this phenomenon every day, as well as more small-scale actions such as browsing a book store or having a conversation. The average modern working day consists of dealing with 46 phone calls, 15 internal memos, 19 items of external post and 22 e-mails [Pat00].

We live in an era of information explosion, with information being measured in exabytes (10^{18} bytes):

> "Print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. (...) We estimate that new stored information grew about 30% a year between 1999

> and 2002. (...) Information flows through electronic channels – telephone, radio, TV, and the Internet – contained almost 18 exabytes of new information in 2002, three and a half times more than is recorded in storage media. (...) The World Wide Web contains about 170 terabytes of information on its surface." [LV03]

On the dawn of the World Wide Web, finding information was done mainly by scanning through lists of links collected and sorted by humans according to some criteria. Automated Web search engines were not needed when Web pages were counted only by thousands, and most directories of the Web included a prominent button to "add a new Web page". Web site administrators were encouraged to submit their sites. Today, URLs of new pages are no longer a scarce resource, as there are thousands of millions of Web pages.

The main problem search engines have to deal with is the size and rate of change of the Web, with no search engine indexing more than one third of the publicly available Web [LG98]. As the number of pages grows, it will be increasingly important to focus on the most "valuable" pages, as no search engine will be able of indexing the complete Web. Moreover, in this thesis we state that the number of Web pages is essentially infinite, which makes this area even more relevant.

1.1.3 Information retrieval and Web search

Information Retrieval (IR) is the area of computer science concerned with retrieving information about a subject from a collection of data objects. This is not the same as Data Retrieval, which in the context of documents consists mainly in determining which documents of a collection contain the keywords of a user query. Information Retrieval deals with satisfying a user need:

> "... the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This 'interpretation' of a document content involves extracting syntactic and semantic information from the document text ..." [BYRN00]

Although there was an important body of Information Retrieval techniques published before the invention of the World Wide Web, there are unique characteristics of the Web that made them unsuitable or insufficient. A survey by Arasu *et al.* [ACGM+01] on searching the Web notes that:

> "IR algorithms were developed for relatively small and coherent collections such as newspaper articles or book catalogs in a (physical) library. The Web, on the other hand, is massive, much less coherent, changes more rapidly, and is spread over geographically distributed computers ..." [ACGM+01]

This idea is also present in a survey about Web search by Brooks [Bro03], which states that a distinction could be made between the "closed Web", which comprises high-quality controlled collections on which a

search engine can fully trust, and the "open Web", which includes the vast majority of Web pages and on which traditional IR techniques concepts and methods are challenged.

One of the main challenges the open Web poses to search engines is "search engine spamming", i.e.: malicious attempts to get an undeserved high ranking in the results. This has created a whole branch of Information Retrieval called "adversarial IR", which is related to retrieving information from collections in which a subset of the collection has been manipulated to influence the algorithms. For instance, the vector space model for documents [Sal71], and the TF-IDF similarity measure [SB88] are useful for identifying which documents in a collection are relevant in terms of a set of keywords provided by the user. However, this scheme can be easily defeated in the "open Web" by just adding frequently-asked query terms to Web pages.

A solution to this problem is to use the hypertext structure of the Web, using links between pages as citations are used in academic literature to find the most important papers in an area. Link analysis, which is often not posible in traditional information repositories but is quite natural on the Web, can be used to exploit links and extract useful information from them, but this has to be done carefully, as in the case of Pagerank:

> "Unlike academic papers which are scrupulously reviewed, web pages proliferate free of quality control or publishing costs. With a simple program, huge numbers of pages can be created easily, artificially inflating citation counts. Because the Web environment contains profit seeking ventures, attention getting strategies evolve in response to search engine algorithms. For this reason, any evaluation strategy which counts replicable features of web pages is prone to manipulation" [PBMW98].

The low cost of publishing in the "open Web" is a key part of its success, but implies that searching information on the Web will always be inherently more difficult that searching information in traditional, closed repositories.

1.1.4 Web search and Web crawling

The typical design of search engines is a "cascade", in which a Web crawler creates a collection which is indexed and searched. Most of the designs of search engines consider the Web crawler as just a first stage in Web search, with little feedback from the ranking algorithms to the crawling process. This is a cascade model, in which operations are executed in strict order: first crawling, then indexing, and then searching.

Our approach is to provide the crawler with access to all the information about the collection to guide the crawling process effectively. This can be taken one step further, as there are tools available for dealing with all the possible interactions between the modules of a search engine, as shown in Figure 1.1.

The indexing module can help the Web crawler by providing information about the ranking of pages, so the crawler can be more selective and try to collect important pages first. The searching process, through

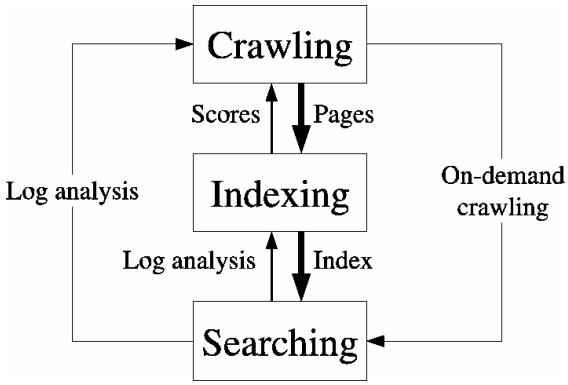

Figure 1.1: Cyclic architecture for search engines. The typical cascade model is depicted with thick arrows.

log file analysis or other techniques, is a source of optimizations for the index, and can also help the crawler by determining the "active set" of pages which are actually seen by users. Finally, the Web crawler could provide on-demand crawling services for search engines. All of these interactions are possible if we conceive the search engine as a whole from the very beginning.

1.2 The WIRE project

At the Center for Web Research (`http://www.cwr.cl/`) we are developing a software suite for research in Web Information Retrieval, which we have called WIRE (Web Information Retrieval Environment). Our aim is to study the problem of Web search by creating an efficient search engine. Search engines play a key role on the Web. Web search currently generates more than 13% of the traffic to Web sites [Sta03]. Furthermore, 40% of the users arriving to a Web site for the first time are following a link from a list of search results [Nie03].

The WIRE software suite generated several sub-projects, including some of the modules depicted in Figure 1.2.

During this thesis, the following parts of the WIRE project were developed:

- An efficient general-purpose Web crawler.
- A format for storing a Web collection.
- A tool for extracting statistics from the collection and generating reports.

Our objective was to design a crawler that can be used for a collection in the order of millions or tens of millions of documents ($10^6 - 10^7$). This is bigger than most Web sites, but smaller than the complete Web, so we worked mostly with national domains (ccTLDs: country-codes top level domains such as `.cl` or `.gr`).

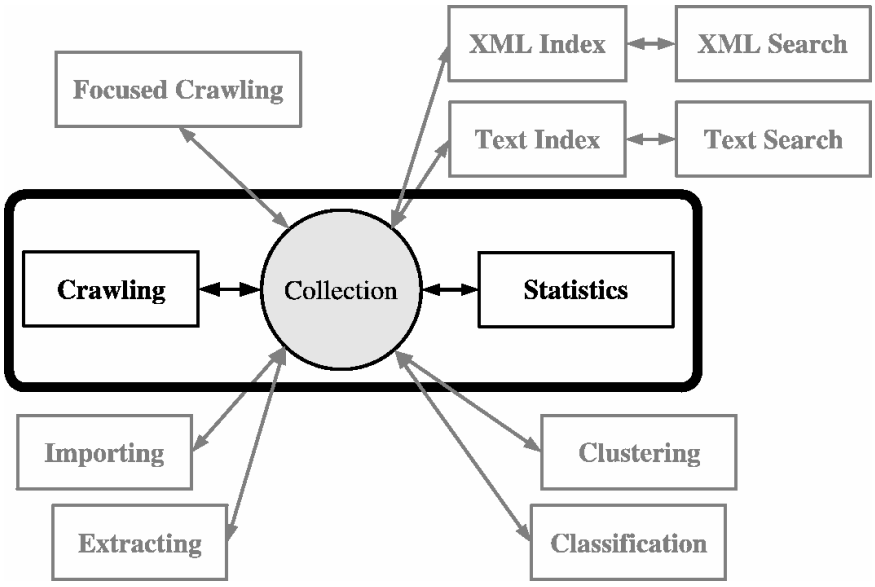

Figure 1.2: Some of the possible sub-projects of WIRE. The thick box encloses the sub-projects in which we worked during this thesis.

The main characteristics of the WIRE crawler are:

Good scalability It is designed to work with large volumes of documents, and tested with several million documents. The current implementation would require further work to scale to billions of documents (e.g.: process some data structures on disk instead of in memory).

Highly configurable All of the parameters for crawling and indexing can be configured, including several scheduling policies.

High performance It is entirely written in C/C++ for high performance. The downloader modules of the WIRE crawler ("harvesters") can be executed in several machines.

Open-source The programs and the code are freely available under the GPL license.

1.3 Scope and organization of this thesis

This thesis focuses on Web crawling, and we study Web crawling at many different levels. Our starting point is a crawling model, and in this framework we develop algorithms for a Web crawler. We aim at designing an efficient Web crawling architecture, and developing a scheduling policy to download pages from the Web that is able to download the most "valuable" pages early during a crawling process.

The topics covered in this thesis are shown in Figure 1.3. The topics are entangled, i.e., there are several relationships that make the development non-linear. The crawler implementation is required for Web

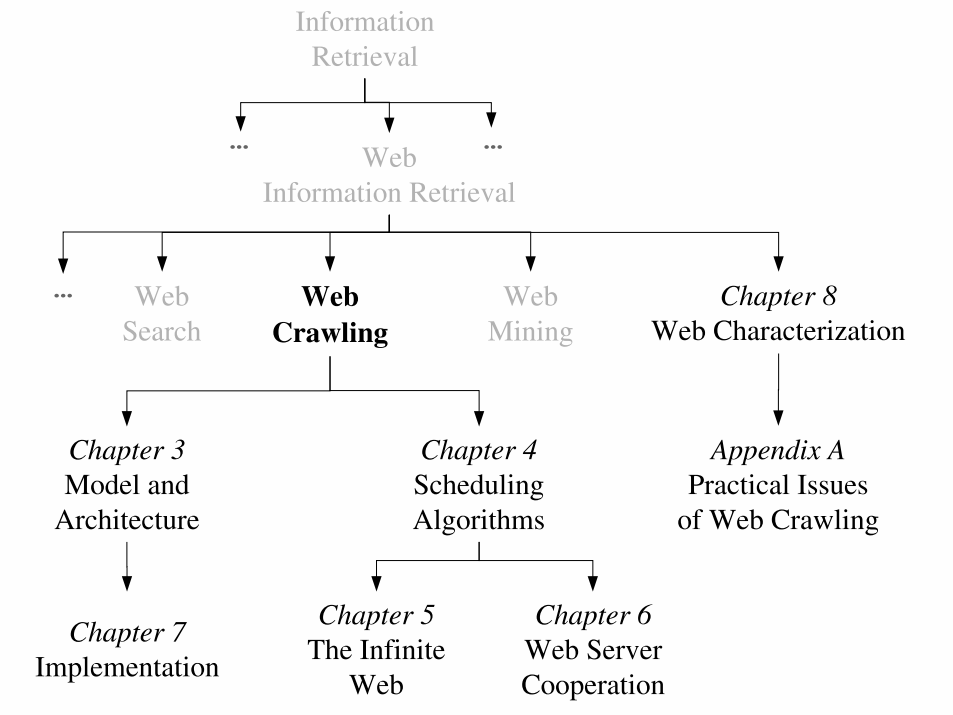

Figure 1.3: Main topics covered in this thesis. Web crawling is important in the context of Web information retrieval, because it is required for both Web search and Web characterization.

characterization, but a good crawler design needs to consider the characteristics of the collection. Also, the crawler architecture is required for implementing the scheduling algorithms, but the result of the scheduling experiments drives the design of the crawler's architecture.

We try to linearize this research process to present it in terms of chapters, but this is not the way the actual research was carried out: it was much more cyclic and iterative than the way it is presented here.

The following is an outline of the contents of this thesis. The first chapters explore theoretical aspects of Web crawling:

- Chapter **??** reviews selected publications related to the topics covered in this thesis, including Web search, link analysis and Web crawling. The next chapters are organized into two parts: one theoretical and one practical.

- Chapter **??** introduces a new model for Web crawling, and a novel design for a Web crawler that integrates it with the other parts of a search engine. Several issues of the typical crawling architectures are discussed and the architecture of the crawler is presented as a solution to some of those problems.

- Chapter **??** compares different policies for scheduling visits to Web pages in a Web crawler. These algorithms are tested on a simulated Web graph, and compared in terms of how soon they are able to find pages with the larger values of Pagerank. We show how in practice we can reach 80% of the total

Pagerank value downloading just 50% of the Web pages.

- Chapter **??** studies an important problem of Web crawling, namely, the fact that the number of Web pages in a Web site can be potentially infinite. We use observations from actual Web sites to model user browsing behavior and to predict how "deep" we must explore Web sites to download a large fraction of the pages that are actually visited.
- Chapter **??** proposes several schemes for Web server cooperation. The administrator of a Web site has incentives to improve the representation of the Web site in search engines, and this chapter describes how to accomplish this goal by helping the Web crawler.

The last chapters empirically explore the problems of Web crawling:

- Chapter **??** details implementation issues related to the design and to algorithms presented in the previous chapters, including the data structures and key algorithms used.
- Chapter **??** presents the results of a characterization study of the Chilean Web, providing insights that are valuable for Web crawler design.

Finally, Chapter **??** summarizes our contributions and provides guidelines for future work in this area.

We have also included a list of issues arising when performing large crawls and more details about the WIRE crawler in the appendices:

- Appendix **??** discusses practical issues of Web crawling that were detected only after carrying several large crawls. We propose solutions for each problem to help other crawler designers.
- Appendix **??** is a copy of the default configuration file for the Web crawler. It is included because it provides a quick overview of its current capabilities.
- Appendix **??** is a copy of the current documentation for the WIRE crawler. It is included because it provides technical information on how the crawler is used in practice.

Finally, the bibliography includes over 150 references to publications in this area. The next chapter is a survey about the most important ones in the context of this thesis.

Bibliography

[ACGM+01] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43, August 2001.

[Bro03] Terrence A. Brooks. Web search: how the Web has changed information retrieval. *Information Research*, 8(3):(paper no. 154), April 2003.

[BYRN00] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison Wesley, 2000.

[Eco03] Umberto Eco. Vegetal and mineral memory: The future of books. http://weekly.ahram.org.eg/-2003/665/bo3.htm, 2003.

[LG98] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[LV03] Peter Lyman and Hal R. Varian. How much information. http://www.sims.berkeley.edu/how-much-info-2003, 2003.

[Nie03] Jakob Nielsen. Statistics for traffic referred by search engines and navigation directories to useit. http://www.useit.com/about/searchreferrals.html, 2003.

[Pat00] Nick Paton. Information overload. *The Guardian*, 2000. (Gallup/Institute for the Future study).

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The Pagerank citation algorithm: bringing order to the web. In *Proceedings of the seventh conference on World Wide Web*, Brisbane, Australia, April 1998.

[PlaBC] Plato. *Phaedrus*. 360 BC.

[Sal71] Gerard Salton. *The SMART retrieval system - experiments in automatic document processing*. Prentice-Hall, 1971.

[SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523, 1988.

[Sta03] StatMarket. Search engine referrals nearly double worldwide. http://websidestory.com/-pressroom/pressreleases.html?id=181, 2003.

[wik04] Library of Alexandria. http://en.wikipedia.org/wiki/Library_of_Alexandria, 2004. (Article on Wikipedia).