Agents, Crawlers, and Web Retrieval*

Ricardo Baeza-Yates and José Miguel Piquer

Center for Web Research
Dept. of Computer Science, University of Chile
Banco Encalada 2120, Santiago, Chile
(rbaeza,jpiquer)@dcc.uchile.cl

Abstract. In this paper we survey crawlers, a specific type of agents used by search engines. We also explore the relation with generic agents and how agent technology or variants of it could help to develop search engines that are more effective, efficient, and scalable.

1 Introduction

The Web has become the largest easy available repository of data. Hence, it is natural to extract information from it and Web search engines have become one of the most used tools in Internet. However, the exponential growth and the fast pace of change of the Web, makes really hard to visit a significant portion of the Web, and even harder to revisit that portion to keep a collection that accurately represents the changes in Web pages, and hence retrieve relevant documents. Therefore, Web retrieval is a practical and interesting research challenge.

Web search engines have three standard components: crawlers, spiders or robots (input), Web database and index (storage) and query resolver and interface (output) [3, chapter 13]. We define a Web database as a collection of Web objects (HTML pages, images, etc.). Current architectures are centralized, with all components running together using parallel processing (typically a farm of PCs).

The main problem of this architecture is scalability, and crawlers are its bottleneck. Crawlers cannot cope with exponential growth and then this decreases Web coverage and Web freshness (having current page content). Another problem is the paradigm used to obtain the input. It is well known that interruptions are better than polling when dealing with slower devices. In practice, Web servers are slower devices, so instead of pulling information from them (only if it has changed or it is new), would be much better that they push the information to the search engine. This idea is explored further later.

We first survey crawlers and we present a taxonomy based in three coordinates. Then we briefly survey agents, and in particular agents as searchers for specific information. We continue with the application of agents and/or pushing technology to improve Web search engines. Finally, we end with some concluding remarks. Related surveys are [2] for Web search engines including crawling and [14] for focused crawling.

* Funded by Millennium Nucleus Center for Web Research, Midepku, Chile.

© Springer-Verlag Berlin Heidelberg 2002
2 Crawlers

All search engines available on the Internet need to traverse Web pages and their associated links, to copy them and to index them into a Web database. The programs associated with the page traversal and recovery are called crawlers. The main decisions associated with the crawlers algorithms are when to visit a site again (to see if a page has changed) and when to visit new sites that have been discovered through links. The parameters to balance are network usage and Web server overload against index accuracy and completeness.

As Web crawlers are mostly used by search engines, their inner components are usually well guarded secrets and only benchmarks [51] are revealed. Some exceptions (in chronological order) are: RBSE Spider [26], Internet Archive [8], SPHINX [40], Google Web crawler [44], Mercator [29], and a parallel crawler [17]. Other papers deal with page ordering for crawling [21, 42], mirroring [16], Web server overloading [55, 54], keep the collection up to date [22, 20, 18], dynamic pages [49], and knowing how Web pages change over time [19, 7, 24]. Recent work focuses in high performance, scalability, and flexibility [43, 5, 25, 15, 52, 45].

The main goals of a crawler are the following [11]:

1. The index should contain a large number of Web objects that are interesting to the search engine’s users.
2. Every object on the index should accurately represent a real object on the Web (content through time).
3. Generate a representation of the objects that capture the most significant aspects of the crawled object using the minimum amount of resources.

We can model different crawlers using three coordinates that are Web page attributes[11]: quality, quantity, and freshness.

Quality refers to the intrinsic semantic value of the object. Quality can be estimated in many ways [21]: link analysis such as Pagerank [44] (link popularity), similarity to a driven query (focused crawling or a searching agent), accesses to that page on the index (usage popularity), and location-based: by the perceived depth (eg. number of directories on the path to the Web object) or by domain name, IP address or geography.

Quantity depends on the size and format of the information being stored for every object. This depends on the representation chosen. If the complete object is stored, it is high. If only a few keywords are stored, it is low.

Freshness relates to the accuracy that the stored representation reflects the current Web page content, because as Web updates are common, freshness decreases fast with time. Freshness can be estimated quite precisely if the last modification date of the Web page is informed by the Web server [18], using the work in [7].

The proposed taxonomy covers many particular cases, as shown in Figure 1.

It must be noticed that the goals mentioned before compete between them, because the crawler must decide between going for a new page, not currently on the index, or refreshing some page that is probably outdated in the index. There