The Observable Web

Yacine Boufkhad, Laurent Viennot

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The Observable Web

Yacine Boufkhad and Laurent Viennot

N° 4790

April 2003

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4790--FR+ENG

INRIA
ROCQUENCOURT

The Observable Web

Yacine Boufkhad* and Laurent Viennot†

Thème 1 — Réseaux et systèmes
Projet Gyroweb

Rapport de recherche n° 4790 — April 2003 — 5 pages

Abstract: The web is now de facto the first place to publish data. However, retrieving the whole database represented by the web appears almost impossible. Some parts are known to be hard to discover automatically, giving rise to the so called hidden or invisible web. On the one hand, search engines try to index most of the web. Almost all related work is based on discovering the web by crawling. This paper is devoted to estimate how accurate is the view of the web obtained by crawling. Our approach is to compare crawling to other ways of discovering the web (mainly by analyzing server or proxy logs of web surfers activity). This work is a first step towards identifying the observable web.

Key-words: Crawl, access logs, hidden web

* Gyroweb – LIAFA, Université Denis Diderot, Case 7014, 2, place Jussieu, F-75251 Paris Cedex 05, boufkhad@liafa.jussieu.fr
† Gyroweb – INRIA Rocquencourt, F-78153 Le Chesnay, Laurent.Viennot@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Le web observable

Résumé : Le web est devenu un médium de premier plan pour publier des données. Cependant, récupérer la totalité de la base de données qu'il représente presque impossible. Certaines parties sont difficiles à découvrir de manière automatique, donnant naissance à ce que l'on appelle le web invisible. D'un autre côté, les moteur de recherche indexent une grande partie du web. La plupart des travaux sur le sujet se basent sur l'exploration du web par la le « crawl », c'est-à-dire en parcourant le graphe du web. Notre approche consiste à comparer la découverte par crawl et d'autres alternatives dont principalement l'analyse des fichiers de logs de serveurs web ou de proxies. Ce travail constitue un premier pas vers la l'identification du web observable.

Mots-clés : Exploration du web (crawl), fichiers de log, web invisible

1 Introduction

Starting from a base set of seed URLs, a crawl [4] consists in retrieving the associated web pages to collect the URLs contained in them and iteratively find new pages.

Following a physicist approach, we consider the web as an intractable object that we try to observe. Crawling is a possible way of operating. Note that various observations can be obtained by changing the crawling parameters. For example, the base set of seed URLs is of utmost importance. Another way of observing the web consists in tracking web surfers activity. This type of information can be found in log files at proxy servers or web servers. Inspecting proxy logs will be relevant for the population of surfers using the proxy. Using web server log will be relevant of the accessed data on the server. Again various observations are obtained depending on the type of log files inspected. A third alternative consists in listing all the files accessible through the HTTP protocol on a given server, which is relevant of the data accessible on that server. Notice that this observation is also partial since dynamic pages will be completely missed here when the two other methods allow to discover some of them.

Related work includes the study of the web structure (namely the web graph) [2] by crawling. Other work state the problem of efficient crawling [3, 1] in terms of URL ordering (which URL should be retrieved first). The main related work studies the possibility of replacing search engines crawlers by indexing directly HTTP traffic in routers [5]. The main goal is then to reduce the traffic due to crawlers. We will see that this approach is not satisfying for indexing the web since many interesting pages may be missed compared to the crawling approach.

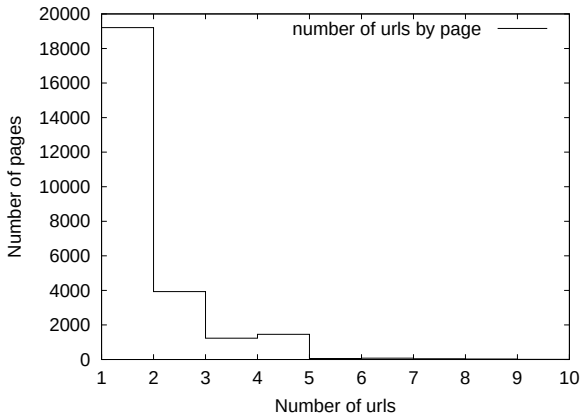

Figure 1: The number of pages found to be associated to a given number of urls.

2 Method

To compare pages that are known by crawling and those accessed by web surfers, we inspect access log files of a server. Hits from robots are separated from browsers hits. On the one hand, we gather together all the hits from the main search engine robots giving a set R of URLs. On the other hand, the set of URLs retrieved by browsers is denoted B.

Notice that search engines benefit from URL submission and start their crawls from very large base sets of URLs. On the other hand, they crawl the whole web and may not crawl very deeply in the server. R is thus further completed with a full local crawl (only internal links are followed). We denote R_c the completed set.

The two sets are then filtered: only html pages are considered and all hits that correspond to outdated pages are discarded. (Practically, we try to retrieve all the URLs of $R_c \cup B$.) Moreover all intranet hits are also discarded, meaning that only pages with no restricted access and allowed by the robots.txt file of the server are considered.

Finally, all pages are identified by a hash of their content. All URLs associated with the same hash are considered to refer to the same page.

Page sets	R_c	B
Pages not in B	1600	0
Pages not in R_c	0	1780
Pages in both B and R_c	13560	13560
Total	15160	15340

Table 1: Pages found by robots and crawl (R_c) versus pages accessed by web browsers (B).

Page sets	$R_c - B$	$B - R_c$
Total size	1560	2040
Personal pages	950	1200
Archives of a forum	334	300
Miscellaneous pages	276	540

Table 2: Content of the differences of the sets R_c and B

3 Experiments

We have performed the above comparison for our own organization server `www-rocq.inria.fr`. The log files of two months of 2002 have been gathered representing three millions of hits. Among the 1.5 millions of hits for html pages, 360,000 came from user browsers and 1.2 millions came from robots. INRIA has a local robot that crawls intensively from the home page of the server. This explains the relatively high rate of robot hits.

The first problem when observing the web consists in identifying the web pages. The only possibility to our knowledge consists in identifying them by their content. This means that two different files with exactly the same content will be considered as the same page with two different URLs. This is acceptable since nothing allows to distinguish them (they include the same information and refer to the same pages). Figure 1 shows the number of pages that have been found to be associated to a given number of urls. URL aliases are often due to symbolic links in the server file system. Some pages may have several thousands of URLs referring to them, this is mainly due to looping symbolic links. This shows that it is important to use a mechanism for identifying uniquely web pages since a quarter of them have more than one URL referring to them.

As `/cgi-bin` was forbidden by the robots.txt file of the server, we have few relevant results about dynamic pages. However, we can still notice significant differences between user hits and robots hits on regular pages.

Table 1 summarizes the sizes of the sets R_c and B of pages found by robots and crawling and pages hit by user browsers respectively. Surprisingly, both sets are comparable in size and agree on almost 90 % of the pages. There are several reasons for that. People usually try to have their pages indexed by robots. This explains why most of pages in B are included in R_c. On the other hand, pages indexed by search engine robots have non-zero probability to appear in response to queries and thus be accessed by search engine users. This is an argument why many pages in R_c are also in B. This fact can be verified by inspecting the referrer field of browser hits. In our experiments, we could identify 70,000 browser hits with a search engine robot as referrer. These hits correspond to 10,000 pages of R_c. We can thus deduce that two third of the pages known by robots have also been accessed by the users of these search engines.

Table 2 gives more insight about the differences between R_c and B. Most of the differences appear on personal pages. Personal pages represent only 15 % of the pages in $R_c \cup B$ and are related to more than 60 % of the differences. On the one hand, some of this pages are not accessible through official link from the server (that would allow to discover the pages when crawling from the home page of the server) and are not advertised to search engine robots. We have no hint why many pages known by robots and not accessed by user browsers are personal pages on the other hand. One hint could be that personal pages are less accessed in general than others, but the inspection of the hits shows that the proportion of hits on personal pages corresponds to the proportion of personal pages.

Following these experiments, we performed a new crawl starting from the set $B - R_c$. Interestingly, we found a set N of 19660 of pages that does neither appear in B nor in R_c. The main part of it is a copy of a large documentation (14,000 pages). However around 5000 new relevant pages (3,000 personal pages and 2,400 miscellaneous pages) were found this way. $R_c \cup B \cup N$ is thus 25 % larger than $R_c \cup B$, showing that both approaches (crawling and logs of user requests) are complementary.

4 Conclusion

Our main conclusion is that the hidden web also contains many regular pages, not only dynamic pages. Another conclusion is that crawling is the best way for finding web pages provided that it starts from a relevant base set of seed URLs. Identifying the observable web can thus be reduced to the problem of gathering this base set. The best way to find or declare such seed URLs is still a challenging task. Inspecting proxy and server log files are a possible source of good seed URLs. However, the legitimate use of log information for indexing purposes may still be discussed.

References

[1] S. Brin, R. Motwani, L. Page, and T. Winograd. The pagerank citation ranking: Bringing order to the web.

[2] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.

[3] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[4] B. Pinkerton. Finding what people want: experiences with the webcrawler. In *Second International WWW Conference*, october 1994. Chicago.

[5] X. Yuan, M. MacGregor, and J. Harms. An efficient scheme to remove crawler traffic from the internet. In *ICCCN'2002*, october 2002. Miami.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399