LD-3

Measuring the Deep and Dark Web

Nobuko Kishi†, Brian Lavoie, Richard Bennett, Edward O'Neill‡

1. INTRODUCTION

Some of the information on the Web is hard to find with search engines. They have been described in various terms such as the invisible Web [1, 2], the dark matter [3] and the deep Web [4]. Although the definitions of these terms differ in their original papers, we adopt three terms: the invisible Web, the dark Web, and the deep Web as follows.

Invisible Web:
The pages whose owners do not want them to be accessible with search engines. They intend to provide private or proprietary information to a limited group of users.

Deep Web:
The pages whose owners want them to be found with search engines, but the search engines choose not to include them in their indexes.

Dark Web:
The pages whose owners want them to be found with the search engines and search engines are interested in including them in their indexes, but the technical difficulties prevent them from doing so.

Table 1 shows the relationship among these classifications, when we call the pages that we can find with search engines Surface Web. In reality the boundaries of these classification are hard to draw, because the technology changes quickly. For example, the information in PDF format used to belong to the Dark Web for a while. However, several search engines have started to index the text part of PDF, putting this information in Surface Web, while other search engines still don't index them and the scanned images in PDF are still in Dark Web.

Table 1. Classification of the information on the Web.

		Search Engine Technique		
		Indexable		Non-Indexable
Owner's Intention	Public	Surface Web	Deep Web	Dark Web
	Private	Invisible Web		

In this paper, we measured the minimum size of these deep, dark and invisible Webs in relative to the surface Web. Our aim is to confirm the existence of the information that we cannot find with search engines. Our long-term aim is to understand the search engines coverage of the information on the Web quantitatively.

We have made the following assumptions to measure the minimum sizes.

1. We can obtain a fair sample of Web sites by sampling IP address space randomly.
2. We can infer the classification of the most of information on the Web from the techniques that used to put pages and links on the Web.

Based on these assumption we made two observations. First, the deep and dark Web has at least the size of about 41 % of surface Web in terms of the number of pages. Second, on the deep and dark Web, the power law holds with the site sizes, but the power law do not hold well with out-degree, the number of out-going links per page, distribution.

2. METHODS

2.0 Method Outline

One of our major assumptions is that the type of the technologies used to provide the information on the Web is closely related to the classifications we adopt. We assume that the following techniques are used in each class of the Web.

- Invisible Web are protected by:
    - Access control by Web server
    - Use of https protocol
    - Use of robots.txt or <meta index >
    - Use of cookies.
- Deep Web is presented by:
    - Use of server-side program for generating pages such as CGI script and servlets.
    - Use of client-side program for generating pages such as JavaScript and Java Applets.
    - Use of input data for the above programs to generate page content, such as <FORM > tags.
- Dark Web consists of
    - Use of non-HTML format data for text. Example: PDF, Word, Power Point file.
    - Use of non-text data. Example: scanned images of printed documents.

Based on these assumptions, we measured the number of pages and links in a sample set of Web server in the following steps.

1. Web server selection by random IP address sampling.
2. Page Collection by two Web crawlers.
3. Classification of collected pages and links into the surface, deep and dark Web.
4. Analysis of page and link distribution.

2.1 Web server selection

The 32-bit IP address space (IPv4) consists of 4,294,967,296 unique IP addresses. A 0.1% random sample (without replacement) was taken from this address space, resulting in 4,294,967 unique IP addresses [5]. An attempt was made to connect to each of the sample IP addresses (on port 80 to locate a Web server and on Port 443 to locate a secure Web server).

When a successful connection was made to a Web server, a request was made to get a root page ("/"). A successful response

† 津田塾大学, Tsuda College
‡ OCLC Online Library Computer Center

code 200 is returned and a request for "/robots.txt" has failed, we consider that IP address hosts a Web server whose information might be either on the surface, deep, or dark Web.

2.2 Crawling Methods

For each Web server we found, we used two different crawlers to collect pages. One, named a limited crawler, is intended to collect the pages on the surface Web, and the other, named an extended crawler, is intended to collect part of the pages on the deep and dark Web. Both of the crawlers follow only relative links to collect the pages on the selected Web server only.

The limited crawler collects the pages whose URLs are in the anchor tags and end with ".htm" or ".html". The extended crawler collects the pages with the following URLs.

1. URLs in ANCHOR tags. URLs may end with any suffix. Example: <a href="test.cgi?name=value&n=v">
2. URLs in the following tags. <area href="xxx"> <frame src="xxx"> <iframe src="xxx"> <link href="xxx"> <meta url="xxx">
3. URLs in the FORM tags. Example: <form action="xxx" >
4. URLs in the VALUE field of OPTION tags. Example: <option value="xxx">
5. Some of URLs in JavaScript. The following parts of HTML text are considered to be JavaScript.
   - o text between <SCRIPT LANGUAGE=javascript> and </SCRIPT >
   - o "onclick", "onselect", "onchange" value of any tags.

   Within these JavaScript texts, the following strings are considered as URLs.
   - o Text strings which ends with ".htm" or ".html"
   - o Text strings which starts with "http://"
   - o Text strings used as the first argument of the "windows.open( )" and "popup()" function calls

2.3 Classification of pages, links and sites

After collecting the pages with the two crawlers, we classify the pages and links, i.e., URLs contained in the collected pages, as follows.

- Pages collected by the limited crawlers are on the surface Web.
- Pages in text format and successfully collected by the extend crawlers are on the deep Web.
- Pages in non-text format and successfully collected by the extend crawlers are on the dark Web.

3. PAGE-BASED SIZE OF DEEP/DARK WEB

We have found 6059 Web servers within the sample IP addresses. The limited crawlers collected 155,920 pages and the extended crawler collected 222,944 pages, about 41% more than the limited crawlers did

Figure 1 shows the log-log plot of the distribution of pages per site. It shows that power law holds for site sizes as reported by Huberman [6]. It also shows a difference between the limited and extended crawlers with large Web servers, which suggests that the deep Web can be found with larger Web servers more often. A drop at the left end of the curve suggests that the power law does not hold well with small Web servers, the servers with less than ten pages.

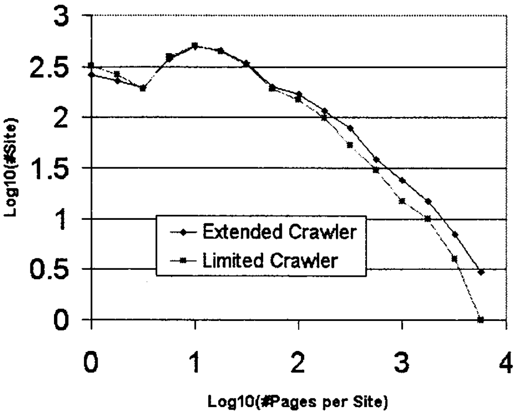

Figure 1. Number of Pages per Site distribution.

4. LINK-BASED SIZE OF DEEP/DARK WEB

Figure 2 shows a distribution of the number of links per page. It shows that the power law holds with the number of links per page, or out-degree distribution as reported by Broder [7]. It also shows the left end of the distribution of all links deviates from the power law, compared to the distribution of off-site (remote-only) links, which agrees with Broder's description.

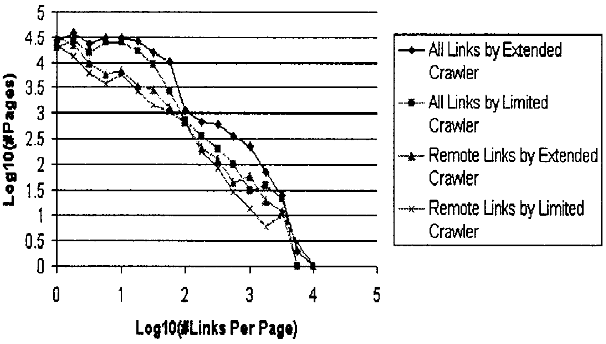

Figure 2: Number of Links per Page Distribution

REFERENCES

[1] C. Sherman and G. Price *The Invisible Web* CyberAge Books. http://www.invisible-web.net

[2] J. Barker *Invisible Web,* http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html

[3] M. K. Bergman The Deep Web: Surfacing Hidden Value, BrightPlanet Corp. 2000.

[4] P. Bailey, N. Craswell and D. Hawking *Dark Matter on the Web,* Poster Proc. 9th WWW, 2000.

[5] E. T. O'Neill, P. D. McClain and Brian F. Lavoie *A Methodology for Sampling the World Wide Web* http://www.oclc.org/research/publications/arr/1997/oneill/o'neillar980213.htm

[6] B. A. Huberman. *The Laws of the Web,* MIT Press. 2001

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener *Graph Structure in the Web,* Proc. 9th WWW, 2000.