DOCUMENT RESUME

ED 459 859 IR 058 411

AUTHOR Darrah, Brenda
TITLE Finding Business Information on the "Invisible Web": Search Utilities vs. Conventional Search Engines.
PUB DATE 2001-05-00
NOTE 22p.; Master of Library and Information Science, Kent State University.
PUB TYPE Dissertations/Theses (040)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Access to Information; Comparative Analysis; Databases; *Information Retrieval; *Internet; Search Strategies; *Small Businesses
IDENTIFIERS *Business Information; *Search Engines

ABSTRACT

Researchers for small businesses, which may have no access to expensive databases or market research reports, must often rely on information found on the Internet, which can be difficult to find. Although current conventional Internet search engines are now able to index over on billion documents, there are many more documents existing in databases that they are unable to reach. A recent study has estimated that the number of these "invisible web" documents may amount to 500 times the number of sites reached by conventional search engines. There are now search utilities that claim to be able to search many of the "invisible" documents. Capabilities in finding business information of the utilities Lexibot and BullsEye 2, as well as Google and metasearch engine Metacrawler, were compared in this study. Searches were conducted for 20 company names, as well as for forecasts, market share, and industry trends for two of these companies. Google was able to find the most relevant results. Because Lexibot and BullsEye 2 did not perform better than general search engines, it appears that it is still necessary to find information from the "invisible web" by first finding appropriate databases, then performing searches. Appendices include a list of companies researched, search terms, and search results data. (Contains 11 references.) (Author/MES)

FINDING BUSINESS INFORMATION ON THE "INVISIBLE WEB": SEARCH UTILITIES VS. CONVENTIONAL SEARCH ENGINES

A Master's Research Paper submitted to the
Kent State University School of Library
and Information Science
in partial fulfillment of the requirements
for the degree Master of Library and Information Science

by

Brenda Darrah

May, 2001

Master's Research Paper by

Brenda Darrah

B.S., University of Akron, 1993

M.L.I.S, Kent State University, 2001

Approved by

Advisor Thomas J. Froehlich Date April 15, 2001

v

CONTENTS

ILLUSTRATIONS

TABLES

INTRODUCTION

Although search engines are currently able to index millions of documents, there are many more contained in the "invisible web." The invisible web contains all the documents buried within databases that search engines are unable to index. In a July 2000 study, BrightPlanet estimated that these documents might amount to 500 times the amount that is indexable by search engines (BrightPlanet, 2000). If this is true, the largest crawler-based search engine, Google (which indexes over 1 billion documents) has only indexed a tiny fraction of the over 500 billion documents on the web (Sullivan, 2000). Moreover, the 60 known, largest invisible web sites contain 40 times more bytes of information than all of the known sites indexable by typical search engines. Some examples of these large invisible sites are the National Climatic Data Center, NASA EOSDIS, the US PTO, and SEC Edgar (BrightPlanet, 2000).

Until recently, the only way to search for information in the invisible web was to first find the individual databases, then search each one. Now there are utilities, such as Lexibot and BullsEye 2, which are able to perform a metasearch of several databases, including invisible ones. Lexibot can perform searches in 600 of the over 22,000 hidden databases listed in BrightPlanet's resource locator tool, CompletePlanet

(BrightPlanet, 2000). It can search up to 60 databases at once (Botluk, 2000). Soon BrightPlanet intends Lexibot to have the capability to search 40,000 web resources, and eventually the estimated 100,000 significant invisible sites (Sullivan, 2000).

BullsEye 2 by Intelliseek supports over 800 search engines (Intelliseek, 2000). BullsEye 2 attempts to ensure relevancy by using a "dynamic intelligent search agent," offering an interface customized to the search (Information Today, 1998).

PROBLEM STATEMENT

In the course of business research, it is often necessary to find information on customers, competitors, or investment prospects that is difficult or expensive to find. If an information seeker has no access to expensive databases or market research reports that can be priced in the thousands of dollars, sometimes free Internet resources can provide solutions. Of course the problem with information on the Internet is finding it, particularly that which is hidden in databases. Tools such as Lexibot and BullsEye 2 are targeted to the research professional searching for such information.

But are these tools really useful? Are the hassle and expense (although less than $100) of downloading, buying, and running these utilities worth the results?

OBJECTIVES

The main objective of this preliminary study was to determine the better of two invisible web search utilities, Lexibot and BullsEye 2. Another objective was to find out whether the invisible web search is really even necessary when compared to conventional search engines or metasearch engines. For this objective, Google and Metacrawler were the sites chosen for comparison because, according to a September 2000 study, they were among the top 5 search sites to provide "the most relevant links in the most logical ranking, with the least effort on our part." (Zetter, 2000)

LITERATURE REVIEW

This paper will focus on relevancy of search results for business research purposes. Over the last few years, there have been several studies ranking search engines based on everything from index size to response time. Only two studies were found that focused on business research, and both tested only conventional web search engines. One of these studies found Google to be the "clear and obvious winner" (Information Advisor, 2000). In the other, Lyle (1999) compared the abilities of seven search engines to find business information

on the Internet. Although he concluded that the Internet could be useful for finding certain business information, particularly product information from company websites, he did not test Google, Metacrawler, or any search utilities like Lexibot and BullsEye 2. No studies were found to evaluate the relevance of results of any type of search on invisible web search utilities.

METHODOLOGY

Testing was conducted in two phases:

Phase I: 20 company names were randomly selected from *Ward's Business Directory of U.S. Public and Private Companies* (see Appendix A for list of names). Each name was entered as a search in all four search engines/utilities. To increase the possibility of finding information, the "Inc.," "Corp.," and "L.L.C." were dropped from the names when performing searches. For example, a query for "BATM Connectronix Corp." in Google returned no results, while a query for "BATM Connectronix" returned seven results, of which five were relevant.

This phase was used to determine whether general information (products, financials, locations, etc.) about these companies could be found using these tools.

Phase II: The following searches were conducted on the two companies tested in Phase I that had the most relevant results:

1) Forecasts for that company's industry (search statement = "industry" forecasts)
2) Information on the industry sales of one of the company's product types (search statement = "product type" revenues)
3) Market share information related to that company (search statement = "company name" market share)
4) Trends in an industry of the company's customers (search statement = trends "customer's industry")

See Appendix B for search terms list.

All searches for each company were performed on the same day. Both BullsEye 2 and Lexibot group search engines by type, such as business, art, education, etc. In this study, the general web search choice was used in both utilities, and all search engines within the general web search group were queried. An attempt to search only business sites in Lexibot resulted in only one result for all 20 companies in Phase I, and many of the business search sites for BullsEye 2 are simply general search engines. So it was decided that the best comparison would be with general web searches. Other than the choice of search engines used, all default settings for each site/utility were used. Company names with more than one word were included in quotes to search for a phrase.

The evaluation version of Lexibot was used. This version has the same capability as the purchased version, except that its use is restricted to 30 days.

There are a few unique features of the search engines/utilities that had to be dealt with to ensure comparability:

- Unlike Lexibot, BullsEye 2, and Metacrawler, Google clusters results by site (Notess, 2000), and the first two results for each site are included in the results. In the results for Lexibot, BullsEye 2, and Metacrawler, any pages after the second of each site were not included when counting the first 25 results.
- A unique feature of Metacrawler is that it imposes time limits on queries sent to each search engine. For this study, when searches timed out, they were tried again with longer timeouts. All of the searches were able to reach all search engines queried when the longer timeouts were used.
- Google provides sponsored links at the top of search results pages. These links were not included in the analysis, as they are not included in Google's results count.

To determine which search engine/utility was best, the number of relevant hits, duplicates, dead links, etc. were analyzed and assigned scores:

2 points: relevant data from a credible source, such as the company's web site, government sites, etc.

1 point: potentially useful and/or provides links for further research. This included sites selling products made by the companies. Results that were links to abstracts from the Northern Light search engine were given one point; although the abstract can be informative, the entire document must be either paid for or found elsewhere. Some results provided useful information, but it was necessary to use information from other websites to determine that it was about the correct company. Some of the companies had been bought out or had changed the name. For example, Mid-Continent Fire and Safety is now Mid-Continent Safety. Because this reflects the reality of web searching, these results were given a score of one.

0 points: not useful, duplicate, or dead link. This included personal home pages and lists of companies that had no indication of: 1) what the list was about, or 2) whether the company is the exact one being researched or not. Dead links are defined as true HTML 404 errors as opposed to temporary page access problems. Any dead links were rechecked after one week

to make sure access problems were not temporary. Duplicate links are defined as links to exactly the same page. This does not include different pages from the same site.

The methodology has been adapted from that of the Information Advisor's study of business searches on "popularity" engines (Information Advisor, 2000). However, this study included searches on more than one company, and the first 25 results were analyzed instead of 10.

DATA ANALYSIS

Phase I Results

Phase I was conducted from January 13 through January 26, 2001. Fig. 1 shows the results. See Appendix C for specific scores for each search.

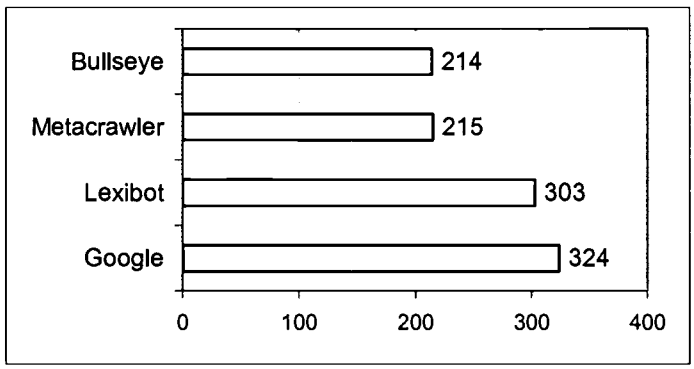

Fig. 1. Phase I Totals

One of the searches conducted in Phase I should be explained. Lexibot returned zero results for "Therm-O-Disc."

When the search was attempted without using dashes, many results were returned. There is no explanation for this. The help function in Lexibot clearly explains that dashes are acceptable in searches.

Phase II Results

Phase II searches were conducted on February 3, 2001 for Independence Blue Cross and North Carolina Eastern Municipal Power Agency, which had the highest scores in Phase I. Fig. 2 shows the results for Phase II. See Appendix D for specific scores for each search.

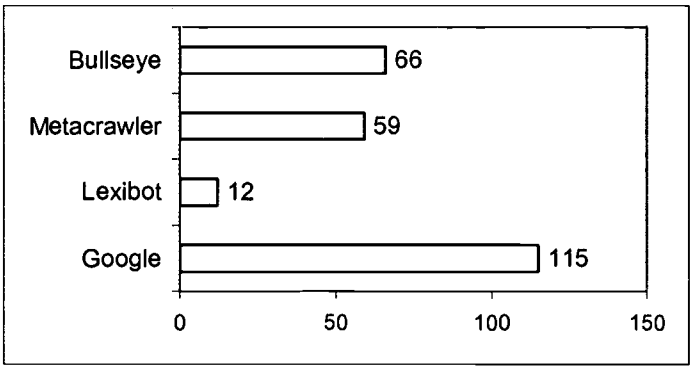

Fig. 2. Phase II Totals

Because electricity customers are the public, the search for electricity customer industry trends was for trends in "electricity use."

Total Results

Fig. 3 shows the combined results of both searches.

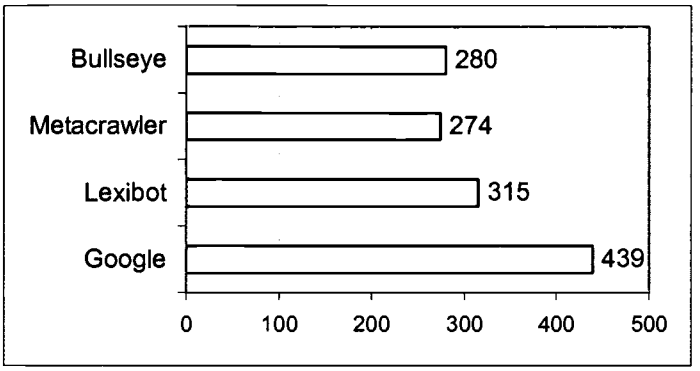

Fig. 3. Total Combined Results

CONCLUSION

It appears that once again, Google is the "clear and obvious winner." Differences in scores in the Phase II results are especially significant. It must be noted that these differences may be due to a change in Google's search capabilities between Phase I and Phase II. Google added the ability to find .pdf files, which often originate from government or educational sites and contain useful information.

Knowing that Lexibot and BullsEye 2 perform no better than general search engines, it seems the best way to find documents from the "invisible web" is still to use directories that link to databases. Google's ability to search .pdf files is very promising, but it seems unlikely that search engines will ever be able to search all of the information on the Internet. It is

still up to the librarian or information professional to point users in the right direction. It would be helpful to use lists of databases, such as InvisibleWeb.com (http://www.invisibleweb.com), Direct Search (http://gwis2.circ.gwu.edu/~gprice/direct.htm), Refdesk.com (http://www.refdesk.com), and Complete Planet (http://www.completeplanet.com). (Botluk, 2000).

APPENDIX A

Companies Researched

American Radar Components, Inc.
39 Front Street
Denville, NJ 07834

BATM Connectronix Corp.
2121 S. 3600 W.
Salt Lake City, UT 84119

Calendar Models of America
P.O. Box 3301
Columbus, OH 43210

Color-Art, Inc.
10300 Watson Road
Sunset Hills, MO 63127

Datamatix, Inc.
215 W. Church Road
King of Prussia, PA 19406

Elite Information Systems, Inc.
5100 W. Goldleaf Cir. #100
Los Angeles, CA 90056

FormTex Plastics Corp.
6817 Wynnwood Ln.
Houston, TX 77008

Hachik Distributors, Inc.
2300 Island Ave.
Philadelphia, PA 19142

Independence Blue Cross
1901 Market St.
Philadelphia, PA 19103

Karen Kane Inc.
2275 E. 37th St.
Los Angeles, CA 90058

Lucas Assembly and Test Systems
12841 Stark Rd.
Livonia, MI 48150

Mid-Continent Fire and Safety, Inc.
2909 S. Spruce St.
Wichita, KS 67216

North Carolina Eastern Municipal Power Agency
P.O. Box 29513
Raleigh, NC 27626

Petro-Global Consultants, Inc.
1900 N. L St.
Midland, TX 79705

Rasch Graphic Services Corp.
8625 Meadowcroft Drive
Houston, TX 77063

Schuyler-Brown FS, Inc.
P.O. Box 230
Rushville, IL 62681

Stanford Ranch I L.L.C.
5146 Arnold Ave.
McClellan AFB, CA 95652

Therm-O-Disc, Inc.
1320 S. Main St.
Mansfield, OH 44907

Vetri Systems, Inc.
2690 Crooks Rd. #305
Troy, MI 48084

Ziegler Ross, Inc.
1 Bay Plaza
Burlingame, CA 94010

APPENDIX B

Phase II Search Terms

- Health insurance industry forecasts
- Health insurance revenues
- Independence Blue Cross market share
- Trends health care
- Electric power industry forecasts
- Electricity revenues
- North Carolina Eastern Municipal Power Agency market share
- Trends electricity use

APPENDIX C

Table 1. Phase I Results

	Google	Lexibot	Metacrawler	BullsEye 2
American Radar Components	1	1	2	1
BATM Connectronix	7	8	12	16
Calendar Models of America	8	9	7	4
Color-Art	6	8	1	10
Datamatix	12	9	8	7
Elite Information Systems	29	29	16	19
FormTex Plastics	12	17	14	12
Hachik Distributors	12	15	13	10
Independence Blue Cross	41	40	20	26
Karen Kane	14	16	11	4
Lucas Assembly and Test Systems	5	10	5	5
Mid-Continent Fire and Safety	8	5	3	7
North Carolina Eastern Municipal Power Agency	43	36	21	18
Petro-Global Consultants	3	2	0	3
Rasch Graphic Services	34	33	23	12
Schuyler-Brown FS	28	29	16	14
Stanford Ranch	13	15	10	10
Therm-O-Disc	26	0	19	22
Vetri Systems	20	21	14	14
Ziegler Ross	2	0	0	0
Total	324	303	215	214

APPENDIX D

Table 2. Phase II Results

	Google	Lexibot	Metacrawler	BullsEye 2
Health insurance industry forecasts	8	2	3	6
Health insurance revenues	4	3	1	2
Independence Blue Cross market share	13	0	3	2
Trends health care industry	23	2	19	23
Electric power industry forecasts	15	1	8	8
Electric power revenues	25	2	16	12
North Carolina Eastern Municipal Power Agency market share	0	0	0	0
Trends electricity use	27	2	9	13
Total	115	12	59	66

WORKS CITED

Botluk, Diana. 2000. *Mining Deeper into the Invisible Web.* Available [Online]: <http://www.llrx.com/features/mining.htm> [9 February 2001].

BrightPlanet. 2000. *The Deep Web: Surfacing Hidden Value.* Available [Online]: <http://www.completeplanet.com/Tutorials/DeepWeb/contents04.asp> [1 November 2000].

"Intelliseek introduces its BullsEye tool for managing information on the Web." *Information Today* 15, no. 11 (December 1998): 23.

Intelliseek. October 2000. Search smarter, not harder. *Intelliseek Newsletter.* Available [Online]: <http://info.intelliseek.com/newsletter/nl_101200.htm> [4 November 2000].

Lyle, S. P. "Comparative study of tools for finding business information on the World Wide Web: a case study." *Journal of Business & Finance Librarianship* 4, no. 2 (Fall 1999): 3-29.

Notess, Greg R. 2000. *Search Engine Statistics: Relative Size Showdown.* Available [Online]: <http://www.searchengineshowdown.com/stats/ size.shtml> [26 October 2000].

"Popularity engines: they're hot, but do they work?" *The Information Advisor* 12, no. 7 (July 2000): 1.

CNET Networks, Inc. 2000. *Search Engine Shootout: How we tested.* Available [Online]: <http://www.cnet.com/internet/0-3817-7-1922954.htm> [14 November 2000].

Sullivan, Danny. 2000. *Invisible Web Gets Deeper.* Available [Online]: <http://searchenginewatch.internet.com/sereport/00/08-deepweb.html> [1 November 2000].

Ward's Business Directory of U.S. Private and Public Companies. Detroit: Gale Research, 2001.

Zetter, Kim; McCracken, Harry; Li-Ron, Yael. "How to stop searching and start finding." *PC World* 18, no. 9 (September 2000): 129-143.

U.S. Department of Education
Office of Educatonal Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").