Search Engine Retrieval of Changing Information

Yang Sok Kim
School of Computing, University of Tasmania, Private Bag 100 Hobart TAS 7001 Australia
yangsokk@utas.edu.au

Byeong Ho Kang
School of Computing, University of Tasmania, Private Bag 100 Hobart TAS 7001 Australia
bhkang@utas.edu.au

Paul Compton
School of Computer Science and Engineering, The University of New South Wales, Sydney 2052 Australia
compton@cse.unsw.edu. au

Hiroshi Motoda
The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Japan
motoda@sanken.osaka-u.ac.jp

ABSTRACT
In this paper we analyze the Web coverage of three search engines, Google, Yahoo and MSN. We conducted a 15 month study collecting 15,770 Web content or information pages linked from 260 Australian federal and local government Web pages. The key feature of this domain is that new information pages are constantly added but the 260 web pages tend to provide links only to the more recently added information pages. Search engines list only some of the information pages and their coverage varies from month to month. Meta-search engines do little to improve coverage of information pages, because the problem is not the size of web coverage, but the frequency with which information is updated. We conclude that organizations such as governments which post important information on the Web cannot rely on all relevant pages being found with conventional search engines, and need to consider other strategies to ensure important information can be found.

Categories and Subject Descriptors
H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval

General Terms
Measurement, Performance, Design

Keywords
Search engine, Web coverage, Overlap of Web search results

1. INTRODUCTION
Although search engine providers have continually competed to expand their coverage, previous research results show that the current coverage of each search engine is significantly different [1-3] and the entire coverage of all search engines is only a fraction of the entire Web [4]. We studied the coverage problem by comparing crawling results with monitoring results assuming that a web monitor would go closer to collecting all the new information pages from given Web information source pages, than a crawler. We compared coverage of the information pages found by our Web monitor program with the coverage of these pages by Google, Yahoo, and MSN. In this paper, we compare the coverage and overlap of three well-known commercial search engines on information pages found by our Web monitor program.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. EVALUATION METHODOLOGY
We selected 260 Australian Government Web pages including both homepages for various departments and media release pages. The Local Government web pages include web pages from both the Tasmanian State Government and municipal government services in Tasmania, thus accounting for the higher number of homepages and smaller number of media release pages compared to the Federal Government. Obviously this sample set will not test the overall performance of Web search engines but we believe that they are not extreme cases with respect to reach-ability by crawlers and frequency of content updating.

Table 1. Sample sites and monitored web pages by domain

Domains		*Web Sites*	*Monitored Pages*
Federal Government	**Homepages**	14 (5%)	1,125 (7%)
	Media release pages	118 (45%)	8,825 (56%)
Local Government	**Homepages**	111 (43%)	2,660 (17%)
	Media release pages	17 (7%)	3,173 (20%)
Total		260	15,770

The Web monitor program, WebMon [5], was used to collect a data set from the sample web pages. At 2 hours intervals, it revisited the Web page to get new information. The monitor identifies new information pages by comparing old URL list (URL_{old}) with new URL list (URL_{new}) of the same monitoring web page and eliminating filtering URLs (URL_{filter}). For each information page the URL, link text, and linked content are stored for further processing and URL_{new} becomes URL_{old}. We collected new Web information pages from August 2005 to October 2006. In total 15,770 new Web pages were collect from the 260 sample Web pages. These are public web pages which should be readily accessible to any web crawler. To check coverage by search engines, we do not simply retrieve the URL as the content may have changed. Rather we submit a query with link text of the collected web page and then check if the page is included amongst those retrieved. We considered 100 search results because 95.5 % of positive results are in the top 100 results with confidence level 95% and confidence interval 5%. Random sampling from the entire data set was necessary in this evaluation because search engines constrain or monitor the

number of automated searches by same user / IP. For each month, we sampled the data set as follows with 95% confidence levels and a 5% confidence interval. 4,203 samples were selected, 23% of all monitoring results.

3. COVERAGE

The overall coverage results for the three search engines are summarized in Table 2. The coverage performance is the proportion of pages or positive result ratio. Google gives the highest overall return and MSN the lowest. Overall Google returns 54% of the monitored pages and MSN 23%. That is they miss from 46% to 77% of the Web information pages that have been posted. The search engines also perform differently across different areas. For Google, local government media release pages give the best results, while local home pages give the worst return. In contrast for both MSN and Yahoo, local government media release pages give the worst results.

Table 2. Coverage Results by Domain

Domains		*Sample*	*Google*	*Yahoo*	*MSN*
Federal	**Home**	289	153(53%)†	87(30%)	106(37%)
	Media	2,328	1,316(57%)	930(40%)	700(30%)
Local	**Home**	724	258(36%)	135(19%)	115(16%)
	Media	862	544(63%)	102(12%)	32(4%)
Total		4,203	2,271(54%)	1,254(30%)	953(23%)

† The ratio is obtained by dividing positive result number with sample page number

Figure 1 illustrates coverage trends during the monitoring period. The month by month results show that Google is consistently the best with Yahoo second, except for an anomalous period at the end, and MSN third. Google and MSN search engines broadly give higher returns in more recent months. This might have been because of improved crawling during the period, but is more likely that they might use crawled date or indexed date as one of results ranking factors. Yahoo does not improve over time, but the sudden change at the end suggests possible changes to the way they crawl the Web.

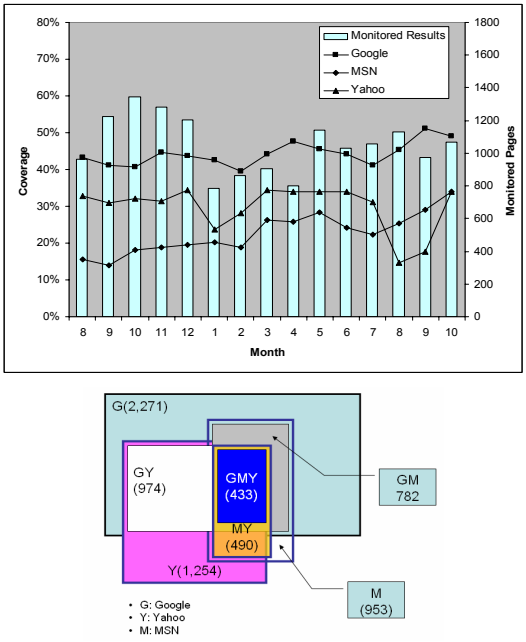

Figure 1. Coverage and Overlap Results

4. OVERLAP AND UNIQUENESS

Total unique positive returns (TUPR) are 2,665, 63.4% of the monitored web pages. It is calculated as follows:

TUPR=G(2,271)+Y(1,254)+M(953)–GM(782)–GY(974)–MY(490)+GMY(433), where G, Y, M, GM, GY, MY, and GMY represent positive results from Google and their overlapped positive returns (see Figure 1 bottom).

Overlap ratio of all search engines is 16.2% (433/2,665) and overlap ratios between pairs of search engines are as follows:

- Google and Yahoo: 974/(2,271(G)+1,254(Y)–974(GY))=38%
- Google and MSN: 782/(2,271(G)+953(M)–782(GM)) = 32%
- MSN and Yahoo: 490/(953(M)+1,254(Y)–490(MY)) = 29%

This result means Google dominates the other search engines because 78% (974/1,254) of Yahoo's positive results are overlapped by Google and 82% (782/953) of MSN's positive results. This result does not suggest a significant improvement by using a meta-search engine.

5. CONCLUSIONS

In this paper we studied coverage and overlap of three commercial search engines (Google, Yahoo, and MSN) using 15,770 Web information pages, which were collected from 260 Australian federal and local government Web pages for 15 months. We found that (1) overall coverage of all three commercial search engines is 63.4% and individually they vary from 22.7% to 54.0%, (2) overall overlap is 16.2 %, which is large compared to other studies [1, 3], and (3) one search engine (Google) is dominant over other search engines, and covers 85% of all unique search returns. We need to enhance coverage by employing dynamic scheduling strategies or use other Web information technologies such as Web monitoring and we need to reconsider the value of meta-search, because our results, especially (2) and (3), weaken the meta-search research assumption of the low coverage of each search engine and low dominance by any one search engine.

6. ACKNOWLEDGMENTS

This work is supported by the Asian Office of Aerospace Research and Development (AOARD) (AOARD-06-4006)

7. REFERENCES

[1] Spink, A., et al., A study of results overlap and uniqueness among major web search engines. Information Processing and Management, 2006. 42(5): p. 1379 - 1391.

[2] Ding, W. and G. Marchionini. A Comparative Study of Web Search Service Performance. in Annual Conference of the American Society for Information Science. 1998.

[3] Bharat, K. and A. Broder. A technique for measuring the relative size and overlap of public Web search engines. in WWW7: The Seventh International World Wide Web Conference. 1998. Brisbane, Australia.

[4] Lawrence, S. and C.L. Giles, Searching the World Wide Web. Science, 1998. 280.

[5] Park, S.S., S.K. Kim, and B.H. Kang. Web Information Management System: Personalization and Generalization. in the IADIS International Conference WWW/Internet 2003. 2003.