The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1066-2243.htm

Overlap among major web search engines

Amanda Spink
Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia
Bernard J. Jansen
School of Information Sciences and Technology, Pennsylvania State University, University Park, Pennsylvania, USA
Vinish Kathuria
Infospace, Inc., Search & Directory, Bellevue, Washington, USA, and
Sherry Koshman
School of Information Sciences, University of Pittsburgh, Pittsburg, Pennsylvania, USA

Abstract

Purpose – This paper reports the findings of a major study examining the overlap among results retrieved by three major web search engines. The goal of the research was to: measure the overlap across three major web search engines on the first results page overlap (i.e. share the same results) and the differences across a wide range of user defined search terms; determine the differences in the first page of search results and their rankings (each web search engine's view of the most relevant content) across single-source web search engines, including both sponsored and non-sponsored results; and measure the degree to which a meta-search web engine, such as Dogpile.com, provides searchers with the most highly ranked search results from three major single source web search engines.

Design/methodology/approach – The authors collected 10,316 random Dogpile.com queries and ran an overlap algorithm using the URL for each result by query. The overlap of first result page search for each query was then summarized across all 10,316 to determine the overall overlap metrics. For a given query, the URL of each result for each engine was retrieved from the database.

Findings – The percent of total results unique retrieved by only one of the three major web search engines was 85 percent, retrieved by two web search engines was 12 percent, and retrieved by all three web search engines was 3 percent. This small level of overlap reflects major differences in web search engines retrieval and ranking results.

Research limitations/implications – This study provides an important contribution to the web research literature. The findings point to the value of meta-search engines in web retrieval to overcome the biases of single search engines.

Practical implications – The results of this research can inform people and organizations that seek to use the web as part of their information seeking efforts, and the design of web search engines.

Originality/value – This research is a large investigation into web search engine overlap using real data from a major web meta-search engine and single web search engines that sheds light on the uniqueness of top results retrieved by web search engines.

Keywords World wide web, Search engines

Paper type Research paper

Internet Research
Vol. 16 No. 4, 2006
pp. 419-426

1066-2243
DOI 10.1108/10662240610690034

1.Introduction

Millions of people use web search engines everyday to find information. Many commercial web search engines offer public access to web sites, including Yahoo!, MSN Search, Google and Teoma. In addition, meta-search engines such as Dogpile.com, Webcrawler, Metacrawler, Hotbot and Mama provide results from multiple single web search engines. Web search engines can differ from one another in three ways – crawling reach, frequency of updates, and relevancy analysis. Therefore, the performance capabilities and limitations of web search engines, and the differences between single and meta-search engines, is an important and significant research area.

There is a critical need for a greater understanding of the differences in web search engines' web site indexing and the overlap among results for the same queries. Research by Ding and Marchionini (1996) first pointed to the often small overlap between results retrieved by different web search engines for the same queries. Lawrence and Giles (1998) also showed that any single web search engines indexes no more than 16 percent of all web sites. These studies began the process of documenting the real differences between web search technologies in terms of indexing, retrieval algorithms and techniques. Our project is part of any ongoing research stream that seeks to understand the characteristics of web search engines and how their content collections are not the same.

The web is also large and millions of new pages are added every day. Gulli and Signorini (2005) estimated the size of the web as 11.5 billion pages. This estimation highlights the difficulty any single web search engine has in attempting to crawl and index the entire web. Moreover, it points to the low likelihood that any single web search engine will have indexed the most recent web pages relevant to a particular query at any one time.

To extend our knowledge of web search engine differences further, we examine the overlap among three major web search engine for results retrieved for the same queries. Our study then compares these three web search engine results with results retrieved for the same queries by the meta-search engine Dogpile.com. Meta-search engines query multiple web search engines concurrently for the same query, combining the results into one listing. Our large-scale study provides an important and significant insight into web search engines differences, and the performance capabilities of single and meta-search engines.

2.Related studies

Since the mid-1990s, web searching research has become a crucial area of study. Jansen *et al.* (2000), Spink *et al.* (2002), and Spink and Jansen (2004) highlight key searching trends from 1997 to 2004, including that most web users do not enter many queries during a search session and view few results pages. Link analysis has also developed as a major web research area (Thelwall, 2004).

Web search engine crawling and retrieving studies are also an important area of web research, including studies that examine the degree of overlap in web search engine results for the same query. Ding and Marchionini (1996), Bharat and Broder (1998), Lawrence and Giles (1998) and Chignell *et al.* (1999) found little overlap in the results returned by various web search engines. Gordon and Pathak (1999) report that approximately 93 percent of results were retrieved by only one web search engine. Nicholson (2000) found similar low web search engine overlap. Gulli and Signorini (2005)

estimated the size of the web as 11.5 billion pages. Cheney and Perry (2005) compare the comparative size of Yahoo! and Google's indexes. Mowshowitz and Kawaguchi (2005) examined the difference between web search engine results from an expected distribution. Egghe and Rousseau (2006) analyze IR system overlap from a mathematical perspective, and Bar-Ilan (2005) discusses a statistical comparison of overlap in web search engines. Bar-Yossef and Gurevich (2006) discuss methods for comparing web search engine indexes.

In summary, studies show that overlap is an important issue for web search engine performance research. Most web search engine overlap studies were performed in the 1990s using small query samples. Given the technological advances since this time in web search engine design, we are examining the current state of web search overlap using a large set of queries. Overall, most web searchers view only the first or second page of results (Spink and Jansen, 2004). Therefore, examining overlap levels for queries on first page results is an important research issue.

3.Research goals

The goals of our research were to:

- Measure the overlap across three major web search engines on the first results page overlap (i.e. share the same results) and the differences across a wide range of user defined search terms.
- Determine the differences in the first page of search results and their rankings (each web search engine's view of the most relevant content) across single-source web search engines, including both sponsored and non-sponsored results.
- Measure the degree to which a meta-search web engine, such as Dogpile.com, provides searchers with the most highly-ranked search results from three major single source web search engines – Google, Yahoo and Ask.com.

4. Research design

4.1 Data collection

To ensure a random and representative sample, the following steps were taken to generate the query list. We pulled 10,316 random queries from the server access log files from Dogpile.com. These key phrases were picked from one weekday and one weekend day of the log files to ensure a diverse set of users. All duplicate queries were removed to ensure a unique list and removed terms that are typically not processed by search engines. We compiled 10,316 random user-entered queries from the Infospace powered network of search site log files.

For each of 10,316 queries in the list, each of the three single web search engines – Google, Yahoo and Ask.com – was queried in sequence between April 17 and 18 of 2005. MSN was under going a re-indexing during this period, so we did not include it in the study. We captured the results (non-sponsored and sponsored) from the first result page and stored the following data in a database: Display URL. Result Position (Note: non-sponsored and sponsored results have unique position rankings because they are separated out on the results page), Result Type (non-sponsored or sponsored).

For non-sponsored results rankings, we looked at main body results that are usually located on the left hand side of the results page. For sponsored result rankings, the study looked at the shaded results at the top of the results page, right-hand boxes usually labeled "Sponsored Results/Links", and the shaded results at the bottom of the

results page for Google and Yahoo!. Ask.com sponsored results are found at the top of the results page in a box labeled "Sponsored web Results".

4.2 Data analysis

After collecting all of the data for the 10,316 queries, we ran an overlap algorithm based on the URL for each result by query. The algorithm was run against each query to determine the overlap of search results by query. When the URL on one engine exactly matched the URL from one or more engines of the other engines a duplicate match was recorded for that query. The overlap of first result page search results for each query was then summarized across all 10,316 to come up with the overall overlap metrics. For a given query, the URL of each result for each engine was retrieved from the database.

A complete result set is compiled for that query in the following fashion: begin with an empty result-set as the complete result set. For each result R in engine E, if the result is not in the complete set yet, add it, and flag that it is contained in engine X. If the result is in the complete set, that means it does not need to be added (it is not unique), so flag the result in the complete set as also being contained by engine X (this assumes that it was already added to the complete set by some other preceding engine).

Determining whether the result is in the complete set or not, is done by simple string comparisons of the URL of the current result and the rest of the results in the complete set. What we have after going through all results for all engines is a complete set of results, where each result in the complete set are marked by at least one engine and up to the maximum number of engines (in this case, three). From this matrix, we can calculate the need metrics to measure overlap.

The research design for this study is further elaborated in the Infospace, Inc white paper by Blakely *et al.* (2005).

5.Results

5.1 First, page overlap results

Table I shows the mean number of results that are unique and shared across the first page results for the three major web search engines.

Overall, a majority of the results a single source web search engine returns on its first result page for a given query are unique to that engine. This data suggests that the differences of each web search engine's indexing and ranking methodologies substantially influence the results a web searcher will receive when searching these engines for the same query. Therefore, while the web search engines in this study may find quality content for some queries, the fact is that they do not always present all of the best content for a given query on their first result page.

Table II shows overall the percent of returned results declines as more web search engines are added to the analysis. Only 3 percent of results were found by all three web search engines.

The number and distribution of sponsored and non-sponsored results on the first page of results is similar across these search engines (i.e. the number and percentages are nearly the same). When looking at sponsored link overlap it makes sense to focus on Yahoo! and Google as they supply sponsored links to the majority of search engines on the web, including MSN Search and Ask.com. Yahoo! returned 37,701 sponsored

	Unique results	Percent	Shared results	Percent
Yahoo!	105,835	31.5		
Shared by Yahoo! and Google			11,209	3.3
Google	86,890	25.8		
Shared by Google and Ask.com			21,201	6.3
Ask.com	93,036	27.7		
Shared by Ask.com and Yahoo!			7,549	2.2
Results found by all three			10,712	3.2
Sub-total	285,761	84.9	50,671	15.1
Total				

Table I. Unique and shared results

	Percentage of the returned results
Unique results to one of the three search engines	85
Unique results to two of the three search engines	12
Unique results to all three search engines	3

Table II. Search engine overlap

links across the 10,316 queries while Google returned 32,774 sponsored links. However, the majority of those were unique to each engine.

The finding also illustrated the known relationships between Google and Ask.com. Through a partnership, Google supplies Ask.com with a feed of their advertisers that Ask.com incorporates into its results page. This partnership is illustrated in the data with a high overlap of sponsored results between Google and Ask.com. Google and Ask.com have a sponsored ink overlap of 14,232 links or 20.6 percent.

Table III shows how the search results ranking differences across the three web search engines for both sponsored and non-sponsored results.

This Table shows that even when the results among search engines overlap, the individual search engines rank the results differently. The ranking for the non-sponsored and sponsored results were measured separately because they are separated on the search results pages.

	Non-sponsored results (percent)	Sponsored results (percent)
Percentage of queries where first result was the same across all three search engines	14.10	3.10
Percentage of queries where top 3 results are the same (not in rank order)	0.14	0.15
Percentage of queries where no URL is the same in any of the top 3 results	31.40	14.90
Percentage of queries where no URL is the same in any of the top 5 results	20.10	11.70

Table III. Ranking for sponsored and non-sponsored results

5.2 Dogpile.com results

Table IV illustrates the results that Dogpile.com displays on its first result page. Dogpile.com.com total first page results for the 10,316 queries were 186,718.

Searching only one web search engine limits a search from finding the best result for their query. Results matched by two or more engines shows the consensus that the results are of value to the query, however, these only account for 15 percent of the total 336,232 links returned on the first results page. Unique results, that represent the largest number of links returned on the first result page of any engine, are valuable when presented with a range of different sources. A meta-search engine such as Dogpile.com presents these unique results from multiple sources.

6.Discussion

This study has produced important findings for all web search engine users, the web industry and researchers. A major result of our study is that first page results returned by the three major web search engines included in this study are different from one another. Web search engines rarely agree on first page returned results for any query. Previous smaller studies have indicated this phenomenon. Despite the advances in search engines technology since the smaller studies were conducted, there is still little agreement among search engines on what is the best results for a given query.

As with all studies, this research has limitations. One concerns the use of only one meta-search engine. Other meta-search engines using other indexes may have different results. In addition, based on our algorithm for comparing URLs, we did not take into account the same destination web site that may use various URLs. Whether this is a serious issue or not is a matter for debate.

Our study of three major commercial web search engines highlight the real differences in web search engines that use different search technologies and produce a high level of uniqueness in sponsored links. Web search engines (Ask.com, Google and Yahoo!) have developed different web indexing and query ranking methods. Meta-search technology, such as Dogpile.com, take the collective content, resources, and ranking capabilities from multiple web search engines to produce a more comprehensive result set containing potentially relevant results from the first results page.

Web search engines continually improve their technology to sort through the growing number of pages in order to return quality results to web searchers. With 26.4 percent of the queries not returning a sponsored link from either Yahoo! or Google, search engine marketers should be aware of the potential missed audience by not leveraging the distribution power of both Google and Yahoo! Those marketers who only optimize for, or purchase on, one web search engine may be missing valuable

Table IV.
Results returned by Dogpile

	Presented by Dogpile (percent)	Not presented by Dogpile (percent)
Results returned by all three engines	95.7	4.3
Results returned by two of the three engines	74.6	25.5
Results returned by only one engine	65.8	34.2

audience exposure by not running on both networks. Therefore, by only running ads on one web search engine limits the coverage. Users need to understand web search engine capabilities, coverage and limitations. Single web search engines have obvious strengths and weaknesses. In some circumstances, the uniqueness of a web search engine's coverage may be useful for engine users.

Our study also has implications for web search engines users. People should know the capabilities, coverage and limitations of the web search engines they seek to use. However, this information is not easy for web users to find.

7. Conclusion and further research

Our study shows that different web search engines have different capabilities and the overlap among web search engine results is very low. The study confirms previous studies and adds new dimensions to our understanding of web searching. These differences contradict the widely held notion that all search engines are the same and that searching one engine will yield the absolute best results of the web. A meta-search engine also provides a unique voice that combines and filters other voices. Further, overlap studies are being conducted using four major web search engines, including MSN Search, to determine additional dimensions of the overlap and rankings.

References

Bar-Ilan, J. (2005), "Comparing rankings of search results on the web", *Information Processing & Management*, Vol. 41, pp. 1511-9.

Bharat, K. and Broder, A. (1998), "A technique for measuring the relative size and overlap of public web search engines", *Computer Networks and ISDN Systems.*, Vol. 30 Nos 1-7, pp. 379-88.

Bar-Yossef, Z.B. and Gurevich, M.G. (2006), "Random sampling from a search engine's index", *Proceedings of the 2006 World Wide Web Conference. 22-26 May 2006. Edinburgh, Scotland*.

Blakely, C., Spink, A. and Jansen, J. (2005), "Different engines, different views web searchers not always finding what they're looking for online", Technical Research Report to Dogpile.com – InfoSpace, Inc, Bellevue, Washington, July.

Cheney, M. and Perry, M. (2005), "A comparison of the size of Yahoo! and Google indices", available at: http://vburton.ncsa.uiuc.edu/indexsize.html

Chignell, M.H., Gwizdka, J. and Bodner, R.C. (1999), "Discriminating meta-search: a framework for evaluation", *Information Processing and Management.*, Vol. 35, pp. 337-62.

Ding, W. and Marchionini, G. (1998), "A comparative study of web search service performance", *Proceedings of the Annual Conference of the American Society for Information Science*, pp. 136-42.

Egghe, L. and Rousseau, R. (2006), "Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve", *Information Processing and Management*, Vol. 42 No. 10, pp. 106-20.

Gordon, M. and Pathak, P. (1999), "Finding information on the world wide web: the retrieval effectiveness of search engines", *Information Processing and Management.*, Vol. 35, pp. 141-80.

Gulli, A. and Signorini, A. (2005), "The indexable web is more than 11.5 billion pages", *Proceedings of the World Wide Web 2005 Conference, May 10-14, Chiba, Japan*.

Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life, real users, and real needs: a study and analysis of user queries on the web", *Information Processing & Management*, Vol. 36 No. 2, pp. 207-27.

Lawrence, S. and Giles, C.L. (1998), "Searching the world wide web", *Science.*, Vol. 280, pp. 98-100.

Mowshowitz, A. and Kawaguchi, A. (2005), "Measuring search engine bias", *Information Processing and Management.*, Vol. 41, pp. 1193-205.

Nicholson, S. (2000), "Raising reliability of web search tool research through replication and chaos theory", *Journal of the American Society for Information Science.*, Vol. 51 No. 8, pp. 724-9.

Spink, A. and Jansen, B.J. (Eds) (2004), *Web Search: Public Searching of the Web*, Springer, Berlin.

Spink, A., Jansen, B.J., Wolfram, D. and Saracevic, T. (2002), *IEEE Computer*, Vol. 35 No. 3, From e-sex to e-commerce: Web search changes, pp. 133-5.

Thelwall, M. (2004), *Link Analysis: An Information Science Perspective*, Elsevier Academic Press.