



## COVER SHEET

---

**Spink, Amanda and Jansen, Bernard J. and Blakely, Chris and Koshman, Sherry (2006) A study of results overlap and uniqueness among major web search engines. *Information Processing and Management* 42(5):pp. 1379-1391.**

**Copyright 2006 Elsevier.**

Accessed from: <http://eprints.qut.edu.au/archive/00004755>

A STUDY OF RESULTS OVERLAP AND UNIQUENESS AMONG MAJOR  
WEB SEARCH ENGINES

Amanda Spink \*  
Faculty of Information Technology  
Queensland University of Technology  
Gardens Point Campus, 2 George St, GPO Box 2434  
Brisbane QLD 4001 Australia  
Tel: 61-7-3864-2782 Fax: 61-7-3864-2703  
Email: [ah.spink@qut.edu.au](mailto:ah.spink@qut.edu.au)

Bernard J. Jansen  
School of Information Sciences and Technology  
The Pennsylvania State University  
University Park PA 16802  
Tel: (814) 865-6459 Fax: (814) 865-6426  
Email: [jjansen@ist.psu.edu](mailto:jjansen@ist.psu.edu)

Chris Blakely  
Market Strategy Manager  
Infospace, Inc. – Search & Directory  
601 108th Ave NE, Ste 1200  
Bellevue, WA 98004 USA  
Tel: (425) 709-8101 Fax: (425) 201-6162  
Email: [chris.blakely@Infospace.com](mailto:chris.blakely@Infospace.com)

Sherry Koshman  
School of Information Sciences  
University of Pittsburgh  
611 IS Building, 135 N. Bellefield Avenue  
Pittsburgh PA 15260  
Tel: (412) 624-9441 Fax: (412) 648-7001  
Email: [skoshman@sis.pitt.edu](mailto:skoshman@sis.pitt.edu)

\* To whom all correspondence should be sent.

## .ABSTRACT

The performance and capabilities of Web search engines is an important and significant area of research. Millions of people world wide use Web search engines very day. This paper reports the results of a major study examining the overlap among results retrieved by multiple Web search engines for a large set of more than 10,000 queries. Previous smaller studies have discussed a lack of overlap in results returned by Web search engines for the same queries. The goal of the current study was to conduct a large-scale study to measure the overlap of search results on the first result page (both non-sponsored and sponsored) across the four most popular Web search engines, at specific points in time using a large number of queries. The Web search engines included in the study were MSN Search, Google, Yahoo! and Ask Jeeves. Our study then compares these results with the first page results retrieved for the same queries by the metasearch engine Dogpile.com. Two sets of randomly selected user-entered queries, one set was 10,316 queries and the other 12,570 queries, from Infospace's Dogpile.com search engine [the first set was from Dogpile, the second was from across the Infospace Network of search properties were submitted to the four single Web search engines. Findings show that the percent of total results unique to only one of the four Web search engines was 84.9%, shared by two of the three Web search engines was 11.4%, shared by three of the Web search engines was 2.6%, and shared by all four Web search engines was 1.1%. This small degree of overlap shows the significant difference in the way major Web search engines retrieve and rank results in response to given queries. Results point to the value of metasearch engines in Web retrieval to overcome the biases of individual search engines.

## INTRODUCTION

Millions of people use Web search engines everyday to find information. Therefore, the performance capabilities and limitations of Web search engines is an important and significant area of investigation. A critical research area is the need for a greater understanding of the differences in Web search engines' Website 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% of all Websites. These studies began the process of documenting the real differences between Web search technologies in terms of indexing, retrieval algorithms and techniques. We are just beginning to understand the characteristics of Web search engines and how their content collections are not the same.

In what ways do Web search engines differ from each other? Currently, we know that Web search engines differ from one another in three primary ways – crawling reach, frequency of updates, and relevancy analysis. The Web is very large and millions of new pages are added every single day. Figure 1 shows the number of textual documents indexed from December 1995 to September 2003 (Search Engine Watch, 2005).

[Place Figure 1 Here]

Today, there are many Web search engine available to Web searchers. ComScore Media Metrix (2005) reported over 166 search engines online in May 2005. Today the indices continue to grow and Table 1 shows where the indices stood as of November 2004 (Search Engine Watch, 2004).

[Place Table 1 Here]

Gulli and Signorini (2005) estimated the size of the Web as 11.5 billion pages. The indices suggest that it is currently difficult for any single Web search engine to crawl and index

the entire Web. Therefore, it is unlikely that all Web search engines will have indexed the most recent Web pages relevant to a particular query at any one time.

To further extend our knowledge of Web search engine differences, this paper reports the results of a major study examining the overlap among four major Web search engine for results retrieved for the same queries. The study then compares these results with the results retrieved for the same queries by the metasearch engine Dogpile.com. Metasearch engines query multiple Web search engines concurrently for the same query, combining the results into one listing. Our study is a significant contribution to Web research as it includes four Web search engines that are the largest search entities operating their own crawling and indexing technology - Ask Jeeves, Google, MSN Search, and Yahoo!. Together, these Web search engines comprise 89.3% of all Web searches conducted in the United States (comScore qSearch Data, April 2005).

Why is the study of Web search engine overlap important? Recent studies by the Pew Internet and American Life Project (2005) show that many people do not understand the capabilities of Web search engines. Some 84.1% of people online use a Web search engine every month to find information (comScore Media Metrix, May 2005). Web searching is also the second most popular online activity, behind email, according to Pew Internet study of Web search engine users (Pew Internet and American Life Project, 2005). Further large-scale studies, such as the one we report, are essential in helping users, Web search companies and researchers understand more about what Web search engines actually accomplish, including the differences between the performance capabilities of single and metasearch engines. Such large-scale studies as ours using commercial Web search engines allow for robust and scalable results often lacking in previous studies.

The next section of the paper situates our study within the previous research investigating Web search engine results overlap.

## RELATED STUDIES

### Overlap Studies

Web research is now a major interdisciplinary area of study, including the modeling of user behavior and Web search engine performance (Spink & Jansen, 2004). Web search engine crawling and retrieving studies have evolved as an important area of Web research since the mid-1990's. In their 1996 study, Ding and Marchionini first identified aspects of the low overlap among the results from the Web search engines InfoSeek, Lycos and Open Text. In a 1996 paper, Gauch, Wang and Gomez also found that a metasearch engine returned the highest number of links judged relevant.

By 1998, Bharat and Broder had measured the size of the Web and overlap between the Websites indexed by the HotBot, Alta Vista, Excite and InfoSeek search engines. They estimated the size of the Web in November 1997 as 200 million pages and the overlap among the Web search engines as 1.4% or 2.2 million pages. Also, in 1998, Lawrence and Giles found that Web search engine coverage of the Web was low and any single Web search engines indexed no more than 16% of all Websites.

In 1999, Chignall, Gwizdka and Bodner found little overlap in the results returned by various Web search engines. Based on their finding, they describe a metasearch engine as useful, since different engines employ different means of matching queries to relevant items and have different indexing coverage. Subsequently, the design and performance of metasearch engines became an ongoing area of study (Buzikashvili, 2002; Chignall, Gwizdka & Bodner, 1999; Dreilinger & Howe, 1997; Meng, Yu & Lui, 2002; Selberg & Etzioni, 1997). Selberg and Etzioni (1999) further suggested that no single search engine is likely to return more than 45% of the relevant results. Gordon and Pathak (1999) studied five search engines and measured overlap at document cut-off values of 20, 50, 100 and 200. They report that approximately 93% of the results were retrieved by only one Web search engine.

Nicholson (2000) replicated the 1996 Ding and Marchionini study and found similar results and low Web search engine overlap. In 2001, Hood and Wilson also found a low overlap amongst bibliographic databases. By 2004, Ferreria, da Silva and Delagardo (2004) stated that studies have shown that documents retrieved by multiple information retrieval (IR) systems in relation to the same query are more likely to be relevant. Mowshowitz and Kawaguchi (2005) examined the difference between Web search engine results from an expected distribution. Egghe and Rousseau (in press) analyze IR system overlap from a mathematical perspective and Bar-Ilan (in press) discusses a statistical comparison of overlap in Web search engines.

While search engine performance studies show little overlap in retrieval, user research has shown that most Web users do not enter many queries and view few results pages (Spink & Jansen, 2004). Click-through studies show few user clicks on Websites (Jansen & Spink, 2003; Mat-Hassan & Levene, 2005). A recent large-scale study conducted by comScore Media Metrix (in press) measured searchers' interaction with first page search results across Ask Jeeves, Google, MSN Search, Yahoo! and Dogpile.com. Between 31 – 55% of searches on the four Web search results and 62.9% of searches on Dogpile.com resulted in a click on the first results page. Clicks on first page results per search ranged from 1.44 to 1.95 for the four single Web search engines and 2.08 for Dogpile.com's first page search result clicks per search. Therefore, any research examining Web search engine overlap should focus initially on the first page of results retrieved as most users are focused on that first page.

In summary, previous studies have produced some consistencies in relation to Web search engine performance, overlap and limitations. These studies highlight differences in Web search engines in terms of Websites indexed and algorithms applied to queries. However, most Web search engine overlap studies were performed in the 1990's using small sets of queries and not targeted at today's major search engines. This paper reports results from large and current study of Web search engine overlap using four major Web search engines – Ask Jeeves, Google, MSN Search, Yahoo! and in comparison to the metasearch engine

Dogpile.com using a large set of queries. The study is a collaboration research project between the Web search industry company Infospace, Inc who provides the meta-search Web search engine Dogpile.com, and academic researchers.

The next section of the paper outlines the study's research design, including the data collection and data analysis.

## RESEARCH GOALS

The goal of our research was to measure the overlap across major Web search engines.

The specific research objectives of the study were to:

- 1) Measure the degree to which the search results on the first results page overlap (i.e., share the same results) as well as differ across a wide range of user queries.
- 2) 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. This analysis includes both sponsored and non-sponsored results.
- 3) Measure the degree to which a metasearch Web engine, such as Dogpile.com, provides searchers with the most highly-ranked search results from each of the four major single source Web search engines.
- 4) Measure any overlap change for the three Web search engines Yahoo!, Google and Ask Jeeves between April and July 2005 (Note: MSN was not included in the April analysis and, therefore, is not included in this section of the study).

The next section of this paper discusses the methodology utilized in this study.

## RESEARCH DESIGN

### Search Result Overlap Methodology

#### Rationale for Measuring the First Result Page

This study set out to measure the first result page of various Web search engines for the following reasons:

- The majority of search result click activity (89.8%) happens on the first page of search results (Infospace internal log files – July 1-14 2005). We view a click as a proxy for interest in a result as it pertained to the search query. Therefore, measuring the first result page captures the majority of activity on search engines.
- Additionally, the first result page represents the top results that an engine found for a given query and therefore is a barometer for the most relevant results an engine has to offer.

#### How the Query Sample Was Generated

To ensure a random and representative sample, the following steps were taken to generate the query list:

1. Pulled random queries (10,316 in April 2005 and 12,570 in July 2005) from the server access log files of the Infospace powered search sites. These key phrases were picked from one weekday and one weekend day of the log files to ensure a diverse set of users.
2. Removed all duplicate queries to ensure a unique list.
3. Removed terms that are typically not processed by search engines.

#### How Search Result Data was Collected

- A. Compiled the two sets of random user-entered queries from the Infospace powered network of search site log files.
- B. Built a tool that automatically queried various search engines, captured the result links from the first result page and stored the data. The tool was a .NET application that queried over http and retrieved the first page of search results. Portions of each result were marked (click URLs) were extracted using regular expressions that were configured per site, normalized, and

stored in a database, along with some information like position of the result and if the result was a sponsored result or not.

C. For each of the queries in the list, each of the four single Web search engines engine was queried between 14-17 April for the 10,316 query set and 15-17 July for the 12,570 query set in sequence (one after another for each query).

- a. Query 1 was run on Ask Jeeves - Google – MSN Search -Yahoo!
- b. Query 2 was run on Ask Jeeves - Google – MSN Search -Yahoo!
- c. Etc.

If an error occurred, the script paused and retried the query until it succeeded. Grabbing the data consisted of making an http request to the site and getting back the raw html of the response. Each query was conducted across all engines within less than 10 seconds. Elapsed time between queries was ~1-2 seconds depending on if an error occurred. The reason for running the data this way was to eliminate the opportunity for changes in indices to impact the data. Each full data set was run in a consecutive 24 - 36 hour window to eliminate the opportunity for changes in indices to impact results.

F. Captured the results (non-sponsored and sponsored) from the first result page and stored the following data in a data base:

- a. Display URL
- b. Result Position (Note: Non-Sponsored and Sponsored results have unique position rankings because they are separated out on the results page)
- c. Result Type (Non-Sponsored or Sponsored)
  - i. For non-sponsored results rankings, we looked at main body results that are usually located on the left hand side of the results page.
  - ii. 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 Jeeves sponsored results are found at the top of the results page in a box labeled 'Sponsored Web Results'.

### How Overlap Was Calculated

After collecting all of the data for the 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.

1. 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.
2. The overlap of first result page search results for each query was then summarized across all queries to come up with the overall overlap metrics.

### Explanation of the Overlap Algorithm

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 X, if the result is not in the COMPLETE set yet, add it, and flag that the result is contained in engine X.
- For each result R 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, 4). The different combinations (in engine X only, in engine Y only, in engine Z only, in both engine X and engine Y but not engine Z, etc.) are then counted up and added to the metric counts being collected for overlap.

The next section of the paper provides the results of our study.

## RESULTS

### First Results Page

#### Mean Number of Results on First Results Page

Table 2 shows the mean number of results that are similar across the first page results for the four major Web search engines for the 12,570 query set.

[Place Table 2 Here]

The mean number of search results returned on the first result page by the four Web search engines is similar as is the proportion of non-sponsored and sponsored results. A mean of 18%-27% of first page search results are sponsored while 73%-82% are non-sponsored. It is important to note that these numbers are averages across the 12,570 queries. The number and distribution of sponsored and non-sponsored results on the first page of results is where the similarity of these engines ends.

#### Search Result Overlap on the First Results Page

Table 3 shows that across the 12,570 queries run on the four engines returned 485,460 unduplicated results.

[Place Table 3 Here]

Of these results:

- 84.9% were unique to one of the four search engines (412,246)
- 11.4% were shared by two of the three search engines (55,515)
- 2.6% were shared by all three search engines (12,398)

- 1.1% were shared by all four search engines (5,301)

These metrics are calculated at the query level and then aggregated. A result like [www.ebay.com](http://www.ebay.com) may appear on multiple engines for various queries. This result is counted as unique each time it shows up on at least one of the engines for a particular query.

#### Missed First Page Web Search Results

Table 4 shows the number and percentage of the possible top results a searcher would have missed had they only used one Web search engine.

[Place Table 4 Here]

Using a single Web search engine only for a query means that a user misses exposure to a range of highly ranked Websites that are provided on the first page of results retrieved to any query. Table 5 below further extends this finding by examining the percentage of first page results that are unique to one Web search engine.

#### Majority of all First Results Page Results are Unique to One Web Search Engine

Table 5 shows the first page results unique to one Web search engine.

[Place Table 5 Here]

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 materially impacts the results a Web searcher will receive when searching these engines for the same query. Therefore, while the engines in this study may find quality content for some queries, the fact is that they do not always find or in some cases present all of the best content for a given query on their first result page.

#### Majority of all First Results Page Non-Sponsored Results are Unique to One Engine

Table 6 shows the percent of first results page non-sponsored results.

[Place Table 6 Here]

Isolating just non-sponsored search results further supports the conclusion that each Web search engine has a different view of the Web. Searching only one Web search engine can limit a searcher from finding the best result for their query.

#### Yahoo! and Google Have a Low Sponsored Link Overlap

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 (i.e., Yahoo!) and Ask Jeeves (i.e., Google).

Table 7 shows the sponsored overlap between Yahoo! and Google.

[Place Table 7 Here]

Yahoo! returned 34,306 sponsored links across the 12,570 queries while Google returned 30,194 sponsored links. However, the majority of those were unique to each engine. The finding also illustrated the known relationships between Google and Ask Jeeves and Yahoo! and MSN Search. Through partnerships, Google supplies Ask Jeeves with a feed of their advertisers that Ask Jeeves incorporates into its results page. Yahoo! supplies MSN Search with a feed of their advertisers that MSN Search incorporates into its results page. These partnerships are illustrated in the data with a high overlap of sponsored results between Google and Ask Jeeves, and Yahoo! and MSN Search.

The sponsored link overlap for these partnerships is:

- Google and Ask Jeeves sponsored link overlap: 14,816 links or 20.6% [I got 25.9%]
- Yahoo! and MSN Search sponsored link overlap: 10,166 links or 17.2% [I got 20.8%]

Analyzing the sponsored links for Yahoo! and Google, the top sponsored link aggregators on the Web, this study found that the number of sponsored links returned was about the only thing these search engines had in common. Yahoo! returned one or more sponsored links for 1,889 queries, which Google did not return any sponsored links. This represents 15% of the total 12,570 queries. Google returned one or more sponsored links for 1,827 queries that Yahoo! did not return any sponsored links. This represents 14.5% of the total

12,570 queries. Almost one third (29.6%) of searches lacked a sponsored result from one of the top sponsored link aggregators.

#### Search Result Ranking Differs Across the Four Search Engines

Table 8 shows how the search results ranking differences across the four Web search engines.

[Place Table 8 Here]

The percentage of the 12,570 queries where the following ranking scenarios were true. Note that non-sponsored and sponsored results were measured separately because they are separated on the search results pages. Ranking matches across all four engines (Ask Jeeves, Google, MSN Search, and Yahoo!).

#### Overlap Comparison Over Time

The comparison of overlap among three of the Web search engines over time (April to July 2005) was examined. Table 9 shows that over time the content on search engines is unique for both sampling periods.

[Place Table 9 Here]

The overlap between Google, Yahoo! and Ask Jeeves fluctuated from April to July 2005 as the percentage of unique results on each of the Web search engines increased slightly.

- The percent of total results unique to one Web search engine grew slightly to 87.7% in July from 84.9% in April.
- The percent of total results duplicated by two Web search engines declined to 9.9% in July from 11.9% in April.
- The percent of total results duplicated by all three Web search engines declined to 2.3% in July from 3.2% in April.

Table 9 shows that across Google, Yahoo!, and Ask Jeeves the percentage change in first page search results slightly more unique in July than April. Both Yahoo! and Google conducted index updates in-between these data runs and the results show they continue to return primarily unique results on the first results page. This data suggests that index updates may affect the content of a search engine and overtime this trend may continue.

#### Dogpile.com Results

Table 10 outlines the results that Dogpile.com displays on its first result page. Dogpile.com total first page results for the 12,570 queries were 231,625.

[Place Table 10 Here]

Table 11 shows that the Dogpile.com total first page non-sponsored results for the 12,570 queries were 145,529.

[Place Table 11 Here]

Table 12 shows that the Dogpile.com total first page sponsored results for the 12,570 queries were 40,786.

[Place Table 12 Here]

Results matched by two or more engines highlight the consensus that the results are of value to the query, however these only account for 15.1% of the total 485,460 links returned on the first results page. Unique results, which represent the largest number of links returned on the first result page of any engine, are useful when presented with an array from different sources thereby mitigating any editorial skew that one engine may have over another (Introna & Nissenbaum, 2000).

## DISCUSSION

This study has produced key findings that are important for all Web search engine users and researchers, and the Web industry. The key finding of our large-scale study is that first results returned by the four major Web search engines included in this study differ from one another. Leading Web search engines rarely agree on which results to return on the first results page for any given search query. This finding confirms previous research results in the up-to-date context of a large study of major commercial Web search engines. The study results highlight the fact that different Web search engines, which use different technology to find and present Web information, yield different first page search results. There is also a high degree of uniqueness in sponsored links between the major paid search providers.

Web search engine's first page results are primarily unique, meaning the other engines did not return the same result on the first result page for a given query. The fact that no one Web search engine covers every page on the Internet and the majority of page one results are unique may contribute to the fact that almost half of all searches on the four major Web search engines fail to elicit a click on a search result. The results also highlight that among Google, Yahoo!, and Ask Jeeves the percentage change in first page search results changed only slightly from April to July 2005. The findings suggest that many Web technical and user related characteristics, such as overlap, number of queries entered, etc. are not dramatically changing over time and further highlights the value of studying Web search trends to gauge the true impact of technological changes.

The results of this study also highlight the fact that the top Web search engines (Ask Jeeves, Google, MSN Search, and Yahoo!), have built and developed proprietary methods for indexing the Web and their ranking of query driven search results differs greatly. Metasearch technology, such as Dogpile.com, harnesses the collective content, resources, and ranking capabilities of all four of the top Web search engines and can deliver Web searchers a more comprehensive result set containing potentially relevant results from the top Web search

engines to the first results page. Since Web content is not static, there are barriers for any one engine's ability to cover the entire Web all of the time. This study suggests that using a metasearch engine that leverages the search power of the top Web search engines may reduce the time spent searching multiple Web search engines while providing the top ranked results from the single Web search engines.

The explosion of information on the Web has created a need for online businesses to continually evolve and remain competitive. To remain competitive, online business, whether an extension of a brick-and-mortar business, a pure-play Internet business, or a content resource, must work to ensure Web searchers can easily find them online. Additionally, Web search engines must continually improve their technology to sort through the growing number of pages in order to return quality results to Web searchers. With 29.6% 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 audience exposure by not running on both networks.

The results suggest that a Web metasearch engine that uses a large number of single Web search engines gives coverage of those sites that each engine has ranked most relevant to the query. According to comScore Media Metrix in a study commissioned by Infospace, 30.5% of Yahoo! searchers, or 19.3 million people, only searched on Yahoo! in January 2005. Similarly, 29.0% percent of Google searchers, or 18.7 million people only searched on Google in January 2005. Therefore, by only running ads on one of these engines a marketer would miss millions of potential customers each month. Metasearch technology that leverages the content of both Google and Yahoo! sponsored listings can effectively bridge this gap. Since sponsored links are relevant for some searches, it is important that end users have the choice to interact with sponsored links when necessary.

A major practical implication for users is – know your Web search engine and know its 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. If they know that metasearch engines are more effective at accessing the top ranked Websites from multiple engine or that a particular search engine focuses on retrieving certain types of Websites (e.g., business, news, homepages, etc.), then that has great value to the user. However, ascertaining this information for Web users is not easy and requires access to good quality information and research about Web search engine capabilities.

#### CONCLUSION AND FURTHER RESEARCH

After 15 years of work, Web search is still in its infancy and technology around Web search will continue to evolve. Our study shows that different Web search engines have different capabilities and the overlap among Web search engine results is very low. The study validates previous studies and adds new dimensions to our understanding of Web searching. Our research conducted to date has uncovered five different voices for Web search based on unique ways of capturing and ranking search results. Google is different than Yahoo! Yahoo! is different from Ask Jeeves. Ask Jeeves is different from MSN Search. These differences contradict the widely held notion that all Web search engines are the same and that searching one engine will yield the absolute best results of the Web. A metasearch engine also provides a unique voice that combines and filters other voices.

Further research is needed to determine additional dimensions of the overlap, across subsequent results pages and rankings of different Web search engines. Additional studies are also necessary to access the strengths and limitations of Web metasearch engines.

## REFERENCES

Bar-Ilan, J. (in press). Comparing rankings of search results on the web *Information Processing and Management*.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1-7), 379-388.

Buzikashvili, N. (2002). Metasearch: Properties of common document distributions. In: U. Karagiannis & U. Reimer (Eds.), *Lecture Notes on Computer Science, Volume 2569* (pp. 226-231) Berlin: Springer.

Chignell, M. H., Gwizdka, J., & Bodner, R. C. (1999). Discriminating meta-search: A framework for evaluation. *Information Processing and Management*, 35, 337-362.

comScore Meta Metrix (in press).

comScore qSearch Data, April 2005.

comScore Meta Metrix, May 2005

Ding, W., & Marchionini, G. (1998). A comparative study of web search service performance. *Proceedings of the Annual Conference of the American Society for Information Science* (pp. 136-142).

Dreilinger, D., & Howe, A. E. (1997). Experiences with selecting search engines using meta-search. *ACM Transactions on Information Systems*, 15(3), 195-222.

Egghe, L. & Rousseau, R. (in press). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. *Information Processing and Management*.

Ferreira, J., da Silva, A. R., & Delgado, J. (2004). Does overlap mean relevance? *Proceedings of WWW/Internet 2004 (IADIS) Conference, Spain, Madrid, October 2004*. LADIS: International Association for Development of the Information Society.

Gauch, S., Wang, G. & Gomez, M. (1996). Profusion: Intelligent fusion from multiple, distributed search engines. *The Journal of Universal Computer Science*, 2(9), 637-649.

Gordon, M. & Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35, 141-180.

Gulli, A., & 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*.

Hood, W. W., & Wilson, C. S. (2001). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology*, 54(12), 1091-1103.

Introna, L., & Nissenbaum, H. (2000). Defining the web: The politics of search engines. *IEEE Computer*, 33(January), 54-62.

Jansen, B. J., & Spink, A. (2003). An analysis of web information seeking and use: Documents retrieved versus documents viewed. *IC'03: Proceedings of the 4th International Conference on Internet Computing, Las Vegas, Nevada, June, 23-26*.

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

Mat-Hassan, M., & Levene, M. (2005). Associating search and navigation behavior through log analysis. *Journal of the American Society for Information Science and Technology*, 56(9), 913-934.

Meng, W., Yu, C., & Lui, K-L. (2002). Building efficient and effective meta-search engines. *ACM Computing Surveys*, 34(1), 48-89.

Mowshowitz, A. & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing and Management*, 41, 193-1205.

Nicholson, S. (2000). Raising reliability of web search tool research through replication and chaos theory. *Journal of the American Society for Information Science*, 51(8), 724-729

Pew Internet and American Life Study. (2005). *Search Engine Users*. Washington D.C.

Search Engine Watch. (2004). <http://www.searchenginewatch.com>.

Selberg, E. & Etzioni, O. 1997. The MetaCrawler architecture for resource aggregation on the Web. *IEEE Expert*, 12(1), 8-14.

Selberg, E., & Etzioni, O. (1999). On the instability of Web search services. *Proceedings of RIAO 2000: Computer-Assisted Information Retrieval, April, Paris, France*.

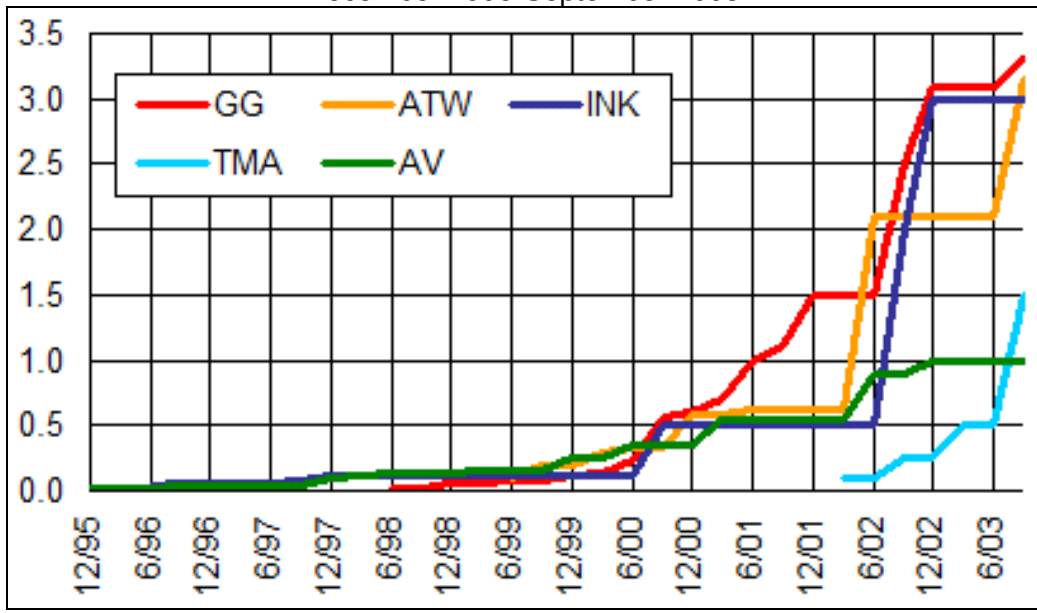
Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. Berlin: Springer:

Appendix A. Examples of random queries.

<b>Keyword</b>	<b>Google Sponsored Links</b>	<b>Yahoo Sponsored links</b>
sunnyside washington death notices for june 26 2002	10	0
kazaa	10	0
native american wedding decorations	10	0
kennel fencing california southern	10	0
outer banks realtors	10	0
berlin flats	10	0
retail fasterns tags	10	0
car graffix	10	0
apartment agencies in berlin	10	0
bulk mail services for real estate	10	0
coleman solar shower	10	0
movers south chicago suburbs	10	0
inexpensive good quality watches	10	0
printing maps for wedding directions	10	0
aluminum ceiling tiles in new jersey	10	0

<b>Keyword</b>	<b>Google Sponsored Links</b>	<b>Yahoo Sponsored Links</b>
washington state university	0	11
peoplepc.com	0	11
knockoff handbags	0	11
replica watches	0	11
fresno wedding services	0	10
land cruiser gx	0	10
mcallaster oklahoma	0	10
carpet cleaning saltash	0	10
louisiana state university	0	10
jacquelina olive oil	0	10
portand maine	0	10
star wars fansite	0	10
tylenol	0	10
bank of america	0	10
era productions	0	10

Figure 1: Billions of textual documents Indexed:  
December 1995-September 2003



**Key:** GG = Google ATW = AllTheWeb INK = Inktomi (now Yahoo!), TMA = Teoma (not Ask Jeeves) AV = Alta Vista (now Yahoo!), Source: Search Engines Watch, January 28, 2005

Table 1: Reported size of each Web search engine index.

Search Engine	Reported Size
Ask Jeeves	2.5 billion
Google	8.1 billion
MSN Search	5.0 billion
Yahoo! (estimate)	4.2 billion

Source: Search Engine Watch, November 11, 2004

Table 2. Mean number of results similar on first results page.

	Total First Page Links	Mean First Page Links Returned	Total Non-Sponsored Links Returned	Mean First Page Non-Sponsored Links Returned	Total Sponsored Links Returned	Mean First Page Sponsored Links Returned
Google	141,973	11.3	111,779	8.9	30,194	2.4
Yahoo!	148,913	11.6	114,607	9.1	34,306	2.7
Ask.com	156,325	12.4	114,497	9.1	41,828	3.3
MSN Search	136,197	10.8	111,398	8.9	24,799	1.9
*Dogpile.com	231,625	18.4	*145,529	*11.6	*40,786	*3.2

\*Note: Dogpile.com's first result page contains results from other Web search engines. These metrics do not take into account the results from other Web search engines not measured in this study.

Table 3. Search result overlap on the first results page.

	Unique	Two Engines	Three Engines	All Four Engines
Google Only	94,293			
Yahoo! Only	106,057			
Ask Jeeves Only	115,525			
MSN Search Only	96,371			
Google & Yahoo!		7,175		
Google & Ask Jeeves		17,279		
Google & MSN Search		7,824		
Yahoo! & Ask Jeeves		5,519		
MSN Search & Yahoo!		14,039		
MSN Search & Ask Jeeves		3,679		
MSN Search & Google		5,336		
Google, Yahoo!, & Ask Jeeves			4,002	
Google, Yahoo!, & MSN Search			3,713	
Yahoo!, Ask Jeeves & MSN Search			2,510	
Google, Ask Jeeves & MSN Search			2,173	
Yahoo!, Google, MSN Search & Ask Jeeves				5,301
Total = 485,460	412,246 (84.9%)	55,515 (11.4%)	12,398 (2.6%)	5,301 (1.1%)

**Table 4.** Number and percentage of the possible top results a searcher would miss using one Web search engine.

	Missed First Page Web Search Results	% of Web's First Page Results Missed
Ask Jeeves	329,761	67.9%
Google	343,700	70.8%
MSN Search	349,561	72.0%
Yahoo!	337,144	69.4%

Table 5. First page results unique to one Web search engine.

	% of Total Results Unique to Search Engine	% of Total Results Overlap with 1+ Search Engines
Ask Jeeves	73.9%	25.7%
Google	66.4%	33.4%
MSN Search	70.8%	29.0%
Yahoo!	71.2%	28.4%

Table 6. Percent of first results page non-sponsored results.

	% of Non-Sponsored Results Unique to Engine	% of Non-Sponsored Results Overlap with 1+ Engines
Google	71.8%	28.2%
Yahoo!	73.9%	26.1%
Ask Jeeves	79.1%	20.6%
MSN Search	73.9%	26.0%

Table 7. Sponsored overlap between Yahoo! and Google.

	Unique Sponsored Links	Overlapping Sponsored Links	% of Engine's Sponsored Links Overlapped
Combined Unique Google & Yahoo! Sponsored Links	61,608	2,892	4.7%

Unduplicated sponsored results between Google and Yahoo! = 61,608

Table 8. Search results ranking differences across the four Web search engines.

	Non-Sponsored Results	Sponsored Results
#1 Result Matched	7.0%	0.9%
Top 3 Results Matched (not in rank order)	0.0%	0.0%
None of Top 3 Results Matched	30.8%	44.5%
None of Top 5 Results Matched	19.2%	41.9%

Table 9. Across Google, Yahoo!, and Ask Jeeves the percentage change in first page search results from April to July 2005.

<b>Overall</b>	<b>April 2005</b>	<b>July 2005</b>
% Unique	84.9%	87.7%
% Overlap with Any Two Engines	11.9%	9.9%
% Overlap with Any Three Engines	3.2%	2.3%
<b>Google</b>		
% Unique	66.7%	71.9%
% Overlap with One Other Engine	24.9%	21.6%
% Overlap with Two Other Engines	8.2%	6.3%
<b>Yahoo!</b>		
% Unique	77.9%	80.6%
% Overlap with One Other Engine	13.8%	12.9%
% Overlap with Two Other Engines	7.9%	6.1%
<b>Ask Jeeves</b>		
% Unique	69.9%	76.3%
% Overlap with One Other Engine	21.6%	17.6%
% Overlap with Two Other Engines	8.0%	5.8%

Table 10. Results that Dogpile.com displays on its first result page.

	% of Dogpile.com Total Results	Total Returned	Total in Dogpile.com
Matched With All 4 Engines	99.3%	5,301	5,264
Matched With Any 3 Engines	95.0%	12,398	11,781
Matched With Any 2 Engines	77.3%	55,515	42,916
Unique to Any One Engine	30.4%	412,246	125,214

Table 11. Dogpile.com.com total first page non-sponsored results.

	% of Dogpile.com Total Results	Total Returned	Total in Dogpile.com
Matched With All 4 Engines	99.5%	4,233	4,213
Matched With Any 3 Engines	96.4%	10,177	9,809
Matched With Any 2 Engines	80.1%	33,212	26,613
Unique to Any One Engine	31.0%	337,923	104,894

Table 12. Dogpile.com total first page sponsored results.

	% of Dogpile.com Total Results	Total Returned	Total in Dogpile.com
Matched With All 4 Engines	98.5%	959	945
Matched With Any 3 Engines	89.3%	2,107	1,881
Matched With Any 2 Engines	73.7%	22,495	16,572
Unique to Any One Engine	28.2%	75,718	21,388