Different Engines, Different Results

Web Searchers Not Always Finding What They're Looking for Online

A Research Study by Dogpile.com

In Collaboration with Researchers from
the University of Pittsburgh and
the Pennsylvania State University

DOGPILE

Executive Summary

In April 2005, Dogpile.com (operated by InfoSpace, Inc.) collaborated with researchers from the University of Pittsburgh (http://www.sis.pitt.edu/~aspink/) and the Pennsylvania State University (http://ist.psu.edu/faculty_pages/jjansen/) to measure the overlap and ranking differences of the leading Web search engines in order to gauge the benefits of using a metasearch engine to search the Web. The study evaluated the search results from 10,316 random user-defined queries across a sample of search sites. The results found that only 3.2% of first page search results were the same across the top three search engines for a given query.

The second phase of this overlap research was conducted in July 2005 by Dogpile.com and researchers from the University of Pennsylvania and the Pennsylvania State University. This study added the recently launched MSN search to the evaluation set of Google, Yahoo! and Ask Jeeves and measured 12,570 user-entered search queries. The results from this latest study highlight the fact there are vast differences between the four most popular single search engines. The overlap across the first page of search results from all four of these search engines was found to be a staggering 1.1% on average for a given query. This paper provides compelling evidence as to why a metasearch engine provides end users with a greater chance of finding the best results on the Web for their topic of interest.

There is a perception among users that all search engines are similar in function, deliver similar results and index all available content on the Web. While the four major search engines evaluated in this study, Google, Yahoo!, MSN and Ask Jeeves do scour significant portions of the Web and provide quality results for most queries, this study clearly supports the last overlap analysis conducted in April 2005. Namely, that each search engine's results are still largely unique. In fact, a separate study conducted in conjunction with comScore Media Metrix found that between 31 – 56% of all searches on the top four search engines are converted to a click on the first result page.[1] With just over half of all Web searches resulting in click-through on the first results page from the top four Web Search Engines at best, there is compelling evidence that Web searchers are not always finding what they are looking for with their search engine.

While Web searchers who use engines like Google, Yahoo!, MSN and Ask Jeeves may not consciously recognize a problem, the fact is that searchers use, on average, 2.8[2] search engines per month. This behavior illustrates a need for a more efficient search solution. Couple this with the fact that a significant percentage of searches fail to elicit a click on a first page search result, and we can infer that people are not necessary finding what they are looking for with one search engine. By visiting multiple search engines, users are essentially metasearching the Web on their own. However, a metasearch solution like Dogpile.com allows them to find more of the best results in one place.

Dogpile.com is a clear leader in the metasearch space. It is highest-trafficked metasearch site on the internet (reaching 8.5 million people worldwide[3]) and is the first and only search engine to leverage the strengths of all the best single source search engines and provide users with the broadest view of the best results on the Web.

To understand how a metasearch engine such as Dogpile.com differentiates from single source Web search engines, researchers from Dogpile.com, the University of Pittsburgh and the Pennsylvania State University set out to:

- Measure the degree to which the search results on the first results page of Google, Yahoo!, MSN, and Ask Jeeves overlapped (were the same) as well as differed across a wide range of user-defined search terms.

- Determine the differences in page one search results and their rankings (each search engine's view of the most relevant content) across the top four single source search engines.

- Measure the degree to which a metasearch engine such as Dogpile.com provided Web searchers with the best search results from the Web measured by returning results that cover both the similar and unique views of each major single source search engines.

Overview of Metasearch

The goal of a metasearch engine is to mitigate the innate differences of single source search engines thereby providing Web searchers with the best search results from the Web's best search engines. Metasearch distills these top results down, giving users the most comprehensive set of search results available on the Web.

Unlike single source search engines, metasearch engines don't crawl the Web themselves to build databases. Instead, they send search queries to several search engines at once. The top results are then displayed together on a single page.

Dogpile.com is the only metasearch engine to incorporate the searching power of the four leading search indices into its search results. In essence, Dogpile.com is leveraging the most comprehensive set of information on the Web to provide Web searchers with the best results to their queries.

Findings Highlight Value of Metasearch

The overlap research conducted in July 2005, which measured the overlap of first page search results from Google, Yahoo!, MSN, and Ask Jeeves, found that only 1.1% of 485,460 first page search results were the same across these Web search engines.

The July overlap study expanded on the April overlap research and measured the recently launched MSN search engine in addition to the previously measured Web search engines. Here's where the combined overlap of Google, Yahoo!, MSN and Ask Jeeves stood as of July 2005:

- The percent of total results unique to one search engine was established to be 84.9%.
- The percent of total results shared by any two search engines was established to be 11.4%.
- The percent of total results shared by three search engines was established to be 2.6%.
- The percent of total results shared by the top four search engines was established to be 1.1%.

Note: Going forward this study will focus on the comparison of all four search engines

Other findings from the study of overlap across Google, Yahoo!, MSN and Ask Jeeves were:

Searching only one Web search engine may impede ability to find what is desired.

- By searching only Google a searcher can miss 70.8% of the Web's best first page search results.
- By searching only Yahoo! a searcher can miss 69.4% of the Web's best first page search results.
- By searching only MSN a searcher can miss 72.0% of the Web's best first page search results.
- By searching only Ask Jeeves a searcher can miss 67.9% of the Web's best first page search results.

Majority of all first results page results across top search engines are unique.

- On average, 66.4% of Google first page search results were unique to Google.
- On average, 71.2% of Yahoo! first page search results were unique to Yahoo!
- On average, 70.8% of MSN first page search results were unique to MSN.
- On average, 73.9% Ask Jeeves first page search results were unique to Ask Jeeves.

Search result ranking differs significantly across major search engines.

- Only 7.0% of the #1 ranked non-sponsored search results where the same across all search engines for a given query.
- The top four search engines do not agree on all three of the top non-sponsored search results as no instances of agreement between all of the top three results were measured in the data.
- Nearly one-third of the time (30.8%) the top search engines completely disagreed on the top three non-sponsored search results.
- One-fifth of the time (19.2%) the top search engines completely disagreed on the top five non-sponsored search results.

Yahoo! and Google have a low sponsored link overlap.

- Only 4.7% of Yahoo! and Google sponsored links overlap for a given query.
- For 15.0 % of all queries Google did not return a sponsored link where Yahoo! returned one or more.
- For 14.5% of all queries Yahoo! did not return a sponsored link where Google returned one or more.

In addition to the overlap results from all four Web search engines, this study measured the overlap of just Google, Yahoo! and Ask Jeeves to compare to the results from the April 2005 study. Findings include:

The overlap of between Google, Yahoo!, and Ask Jeeves fluctuated from April to July 2005. Period over period the percentage of unique results on each of these engines grew slightly.

First page search results from the top Web search engines are largely unique.

- The percent of total results unique to one search engine grew slightly to 87.7% (up from 84.9%).

- The percent of total results shared by any two search engines declined to 9.9%, down from 11.9%.
- The percent of total results shared by three search engines declined to 2.3%, down from 3.2%.

It is noteworthy that both Yahoo! and Google conducted major index updates in-between these studies which most likely effected overlap, a trend that will most likely continue as each engine continues to improve upon their crawling and ranking technologies.

In order to get the best quality search results from across the entire Web, it is important to search multiple engines, a task Dogpile.com makes efficient and easy by searching all the leading engines simultaneously and bringing back the best results from each.

Table of Contents

Introduction

Over the past 18 months, the Web search industry has undergone profound changes. Heavy investment in research and development by the leading Web search engines has greatly improved the quality of results available to searchers. Earlier this year marked the fourth major entry into the search market with the launch of MSN's search index. The rapid growth of the Internet, coupled with the desire of the leading engines to differentiate themselves from one another gives each engine a unique view of the Web causing the results returned by each engine for the same query to differ substantially.

In this study, researchers investigated the difference in search results among four of the most popular Web search engines using 12,570 queries and 485,460 sponsored and non-sponsored results. Results show that overlaps among search engine results are between 25-33% and that less than 20% of the time engines agree on any of the top five ranked search results. These findings have a direct impact on search engine users seeking the best results the Web has to offer. For individuals, it means that no single engine can provide the best results for each of their searches, all of the time.

To quantify the overlap of search results across Google, Yahoo!, Ask Jeeves and MSN Search, we performed the same query at each Web search engine, captured and stored first results page search results from each of these search engines across a random sample of 12,570 user-entered search queries. For this study, a user-entered search query is a full search term/phrase exactly as it was entered by an end-user on any one of the InfoSpace Network powered search properties. Queries where not truncated and the list of 12,570 was de-duplicated so there were no duplicate queries measured.

Background

Today, there are many search engine offerings available to Web searchers. comScore Media Metrix reported 166 search engines online in May 2005[4]. With 84.2%[5] of people online using a search engine to find information, searching is the second most popular activity online according to a Pew Internet study of search engine users (2005)[5].

Search engines differ from one another in two primary ways – their crawling reach and frequency or relevancy analysis (ranking algorithm).

Web Crawling Differences

The Web is infinitely large with millions of new pages added every day.
Statistics from Google.com, Yahoo.com, Cyberatlas and MIT current to April 2005 estimate:

- 45 billion static Web pages are publicly-available on the World Wide Web. Another estimated 5 billion static pages are available within private intranet sites.
- 200+ billion database-driven pages are available as dynamic database reports ("invisible Web" pages).

Estimates from researchers at the Università di Pisa and University of Iowa put the indexed Web at11.5 billion pages[7] with other estimates citing an additional 500+ billion non-indexed and invisible web pages yet to be indexed.[8]

Taking a look back, the amount of the Web that has been indexed since 1995 has changed dramatically.

Billions Of Textual Documents Indexed
December 1995-September 2003

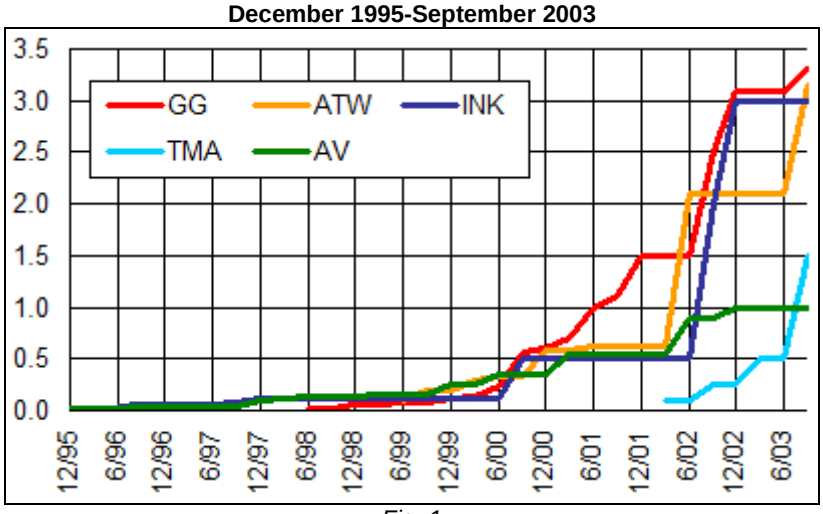

Fig. 1

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

Today, the indices continued to grow. The size of the Web, and the fact that content is ever changing makes it difficult for any search engine to provide the most current information in real- time. In order to maximize the likelihood that a user has access to all the latest information on a given topic, it is important to search multiple engines.

Based on a recent study conducted by A. Gulli and A.Signorini[7] there is a considerable amount of the Web that is not indexed or covered by any one search engine. Their research estimates the visible Web (URLs search engines can reach) to be more than 11.5 billion pages, while the amount that has been indexed to date to be roughly 9.4 billion pages.

Search Engine	Self-Reported Size (Billions)	Estimated Size (Billions)	Coverage of Indexed Web (%)	Coverage of Total Web (%)
Google	8.1	8.0	76.2	69.6
Yahoo!	4.2 (est.)	6.6	69.3	57.4
Ask	2.5	5.3	57.6	46.1
MSN (beta)	5.0	5.1	61.9	44.3
Indexed Web	N/A	9.4	N/A	N/A
Total Web	N/A	11.5	N/A	N/A
Note: “Indexed Web” refers to the part of the Web considered to have been indexed by search engines.				

Fig. 2 Source: A. Gulli & A. Singorini, 2005

Relevancy Differences

Relevancy analysis is an extremely complex issue, and developments in this area represent some of the most significant progress in the industry. The problem with determining relevancy is that two users entering the very same keyword may be looking for very different information. As a result, one engine's determination of relevant information may be directly in the line with a user's intent while another engine's interpretation may be off-target. A goal of any search engine is to maximize the chances of displaying a highly ranked result that matches the users' intent. With known differences in crawling coverage it is necessary for users to query multiple search engines to obtain the best information for their query.

While no search engine can definitively know exactly what every person intends when they search, a searcher's interaction with the results set can help in determining how well an engine does at providing good results. Dogpile.com in conjunction with comScore Media Metrix devised a measure for tracking searcher click actions after a search is entered and quantifying:

- If a searcher clicks one or more search results
- The page which a user clicked a search result
- The volume of clicks on search results generated for each search

A search that results in a click implies that a search result of value was found. Searches that result in a click on the first result page implies the search engine successfully understood what the user was looking for and provided a highly-ranked result of value. Searches that result in multiple clicks imply that the search engine found multiple results of value to the user.

This paper presents the results of a study conducted to quantify the degree to which the top results returned by the leading engines differ from one another as well as how well Dogpile.com's metasearch technology mitigates these differences for Web searchers. The numbers show a striking trend that the top-ranked results returned by Google, Yahoo!, MSN Search and Ask Jeeves are largely unique. This study chose to focus on these four engines because they are the largest search entities that operate their own crawling and indexing technology and together comprise 89.0%[9] of all searches conducted in the United States.

The Parts of a Crawler-Based Search Engine

Crawler-based search engines have three major elements. First is the spider, also called the crawler. The spider visits a Web page, reads it, and then follows links to other pages within the site. This is what is commonly referred to as a site being "spidered" or "crawled". The spider returns to the site on a monthly or bi-monthly basis to look for changes.

Everything the spider finds goes into the second part of the search engine, the index. The index, sometimes called the catalog, is like a giant book containing a copy of every Web page that the spider finds. If a Web page changes, then this index is updated with new information.

Sometimes it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a Web page may have been "spidered" but not yet "indexed." Until it is indexed – (added to the index) -- it is not available to those searching with the search engine.

The third part of a search engine is the search engine software that sifts through the millions of pages recorded in the index to find matches to a search query and rank them in order of what it believes is most relevant.

Major Search Engines: The Same, But Different

All crawler-based search engines have the basic parts described above, but there are differences in how these parts are tuned. This is why the same search on different search engines will often produce dramatically different results. Significant differences between the major crawler-based search engines are summarized on the Search Engine Features Page. Information on this page has been drawn from the help pages of each search engine; along with data gained from articles, reviews, books, independent research, tips from others and additional content received directly from the various search engines.

Source: Search Engine Watch Article, "How Search Engines Work", Danny Sullivan, October 14, 2002.

Search Engine Overlap Studies

Research has previously been done on this topic. Some much smaller studies have suggested the lack of overlap in results returned for the same queries. Web research in 1996 by Ding and Marchionini (1996) first pointed to the often small overlap between results retrieved by different search engines for the same queries. And in 1998, Lawrence and Giles (1998) showed that a single Web search engines indexes no more than 16% of all Web sites.

Search Result Overlap Methodology

Rationale for Measuring the first Result Page

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

- According to Dogpile.com, the majority of search result click activity (89.8%) happens on the first page of search results[10]. For this study a click was used 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 an engine found for a given keyword and is therefore a barometer for the most relevant results an engine has to offer.

How Query Sample was Generated

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

1. Pulled 12,570 random keywords from the Web server access log files from the InfoSpace powered search sites. These key phrases were picked from one weekday and one weekend day of the log files to ensure a more diverse set of users.
2. Removed all duplicate keywords to ensure a unique list
3. Removed non alphanumeric terms that are typically not processed by search engines.

How Search Result Data was Collected

A. Compiled 12,570 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 Google, Yahoo!, Ask Jeeves (Ask.com), and MSN Search over http and retrieved the first page of search results. Portions of each result (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 keyword in the list (the study used 12,570 user entered keywords), each engine of interest (Google, Yahoo!, Ask Jeeves (Ask.com), and MSN Search) was queried in sequence (one after another for each keyword).
   a. Query 1 was ran on Google – Yahoo! - Ask.com – MSN Search
   b. Query 2 was run on Google – Yahoo! - Ask.com – MSN Search, 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. The full data set was run in a consecutive 24-36 hour window to eliminate the opportunity for changes in indices to impact results.

D. 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 the are separated out on the results page)
   c. Result Type (Non-Sponsored or Sponsored)
      i. For Algorithmic results rankings we looked at main body results which are usually located on the left hand side of the results page. See Appendix B.

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.com sponsored results are found at the top of the results page in a box labeled ‘Sponsored Web Results’. See Appendix C.

How Overlap Was Calculated

After collecting all of the data for the 12,570 queries, we ran an overlap algorithm based off the display URL for each result. The algorithm was run against each query to determine the overlap of search results by query.

- When the display URL on one engine exactly matched the display URL from one or more engines of the other engines a duplicate match was recorded for that keyword.
- The overlap of first result page search results for each query was then summarized across all 12,570 queries to come up with the overall overlap metrics.

Explanation of the Overlap Algorithm

For a given keyword, the URL of each result for each engine was retrieved from the database. A COMPLETE result set is compiled for that keyword 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's 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.

The end result 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.

Findings

Average Number of Results Similar on First Results Page

The average number of search results returned on the first result page by the top four engines is similar as is the proportion of non-sponsored and sponsored results.

	Total 1st Page Links	Avg. # 1st Page Links Returned	Total Algorithmic Links Returned	Avg. # 1st Page Algorithmic Links Returned	Total Sponsored Links Returned	Avg. # 1st 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	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

Fig. 3

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

On average 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.

Low Search Result Overlap on the First Results Page Across Google, Yahoo!, MSN Search and Ask Jeeves

Across the 12,570 queries run on Google, Yahoo!, Ask.com and MSN Search, these four engines returned 485,460 unduplicated results. Of these results:

- 1.1% were shared by all four search engines (5,301)
- 2.6% were shared by all three search engines (12,398)
- 11.4% were shared by two of the three search engines (55,515)
- 84.9% were unique to one of the four search engines (412,246)

Note: These metrics are calculated at the query level and then aggregated. Therefore a result like 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 query.

	Unique	Two Engines	Three Engines	All Four Engines
Google Only	94,293			
Yahoo! Only	106,057			
Ask.com Only	115,525			
MSN Search Only	96,371			
Google & Yahoo!		7,175		

	Unique	Two Engines	Three Engines	All Four Engines
Google & Ask.com		17,279		
Google & MSN		7,824		
Yahoo! & Ask.com		5,519		
MSN Search & Yahoo!		14,039		
MSN Search & Ask.com		3,679		
MSN Search & Google		5,336		
Google, Yahoo!, & Ask.com			4,002	
Google, Yahoo!, & MSN			3,713	
Yahoo!, Ask.com, & MSN			2,510	
Google, Ask.com,& MSN			2,173	
Yahoo!, Google, MSN, & Ask.com				5,301

Fig. 4

Searching only one search engine will not yield the best results from the Web all of the time.

Searching Only One Web Search Engine may Impede Ability to Find What is Desired

For this study there were 485,460 unique first page search results across these four Web search engines. The following grid illustrates the number and percentage of the possible top results a searcher would have missed had they only used one Web search engine.

	Missed 1st Page Web Search Results	% of Web's 1st Page Results Missed
Google	343,700	70.8%
Yahoo!	337,144	69.4%
MSN	349,561	72.0%
Ask Jeeves	329,761	67.9%

Fig. 5

Sponsored Link Matching Differs

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 sites had in common.

Yahoo! returned one or more sponsored links for 1,889 keywords which Google did not return any sponsored links. This represents 15.0% of the total 12,570 queries.

Google returned one or more sponsored links for 1,827 keywords which Yahoo! did not return any sponsored links. This represents 14.5% of the total 12,570 queries.

Overall, nearly one-third, 29.6%, of all searches lacked a sponsored result from one of the top sponsored link aggregators.

Majority of all first Results Page Results are Unique to One Engine

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

Fig. 6

Overall, a majority of the results a single source 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 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

	% 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.com	79.1%	20.6%
MSN	73.9%	26.0%

Fig. 7

Isolating just non-sponsored search results further supports the fact that each engine has a different view of the Web. Searching only one search engine can limit a searcher from finding the best result for their query. For those using a search engine to research a topic this data highlights a need to search multiple sources to fully explore a topic whether it is researching ancient Mayan civilization or vacation packages to Hawaii.

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 and Ask.com.

The study found 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.

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

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

Fig. 8

The study also illustrated the known relationships between Google and Ask Jeeves and Yahoo! and MSN. Through partnerships, Google supplies Ask Jeeves with a feed of their advertisers that Ask Jeeves incorporates into its results page. Yahoo! supplies MSN with a feed of their advertisers that MSN 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.

Here's what the sponsored link overlap looks like for these partnerships:

- Google and Ask Jeeves sponsored link overlap: 14,816 links or 20.6%
- Yahoo! and MSN sponsored link overlap: 10,166 links or 17.2%

Search Result Ranking Differs Across Major Search Engines

Figure 9 illustrates 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 (Google, Yahoo!, MSN, and Ask.com)

	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%

Fig. 9

Overlap Composition of First Page Search Results Unique to Each Engine.

The comparison of overlap among engines over time (April 2005 to July 2005) illustrates that over time the content on search engines is unique. 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 will most likely continue.

Data from April's original three engine study (Google, Yahoo! and Ask Jeeves) was compared to data using the same three engines in July 2005. The four data tables shown below on this page refer to comparisons between these three engines only, and do not include MSN data.

Across Google, Yahoo!, and Ask Jeeves the percentage change in unique first page search results was 3.3%. The results from these engines were slightly more unique in July than April.

Overall	April 2005	July 2005
% Unique	84.9%	87.7%
% Overlap with Any Two Engines	11.9%	9.9%
% Overlap with Any Three Engines	3.2%	2.3%

Fig. 10

The percentage change in Google's first page unique search results was 7.8%. Google's first page search results were more unique in July than in April.

Google	April 2005	July 2005
% Unique	66.7%	71.9%
% Overlap with One Other Engine	24.9%	21.6%
% Overlap with Two Other Engines	8.2%	6.3%

Fig. 11

The percentage change in Yahoo's first page unique search results was 3.5%. Yahoo's first page search results were slightly more unique in July than in April.

Yahoo!	April 2005	July 2005
% Unique	77.9%	80.6%
% Overlap with One Other Engine	13.8%	12.9%
% Overlap with Two Other Engines	7.9%	6.1%

Fig. 12

The percentage change in Ask Jeeves' first page unique search results was 9.2%. Ask Jeeves' first page search results were more unique in July than in April.

Ask Jeeves	April 2005	July 2005
% Unique	69.9%	76.3%
% Overlap with One Other Engine	21.6%	17.6%
% Overlap with Two Other Engines	8.0%	5.8%

Fig. 13

Support Research – Success Rate

Since a searcher's actual intent is difficult to quantify, a method was devised to gauge the performance of search engines and their ability to interpret a searcher's intent. By measuring a searcher's interaction with a search engine, specifically their click behavior on search results, insight into relative Web search engine performance can be established. By quantifying if a search resulted in a click on a search result as well as the number of search results clicked we can measure the degree to which an engine's results set provides a satisfactory experience to the searcher.

The study, conducted by comScore Media Metrix and commissioned by InfoSpace, set out to measure the interaction of searchers with first page search results across Google, Yahoo!, MSN, Ask Jeeves. The study found that between 31 – 56% of all searches on these top four search results in a click on a result on the first result page. This measure is called the Success Rate for the search engine. The relatively low Success Rate for the top Web search engines is astonishing and further evidence that Web searchers do not always find what they are looking for with their search engine. However, Dogpile.com's metasearch results converted 62.9% of searches to a click on the first results page.

Additionally, the study measured the differences in click volumes on Web search results. Click volumes speak to the volume of results the searcher found of value to the query conducted. Clicks on first page results per search ranged from 1.36 to 1.95 for the top four single source search engines, while Dogpile.com's metasearch results garnered 2.08 first page search result clicks per search.

When looking at both the percentage of searches that elicit a click (Success Rate), and the number of first page search result clicks per search, metasearch engine Dogpile.com's approach of pulling the top results from the top engines together in one place yields the highest conversion rate of searches to a clicks as well as the most first page search result clicks per search.[1]

	Search to Click Conversion Rate (Success Rate)	% of Searches that fail to elicit a 1st result page click	Clicks per Successful Search
Dogpile.com	62.9%	37.1%	2.08
Google	55.6%	44.4%	1.95
Yahoo!	50.0%	50.0%	1.57
MSN	46.6%	53.4%	1.36
Ask Jeeves	39.7%	68.6%	1.44

Fig. 14

What Metasearch Engine Dogpile.com Covers

The above data has illustrated that there are differences in what the top four single source search engines deem as important results as measured by being returned on the first results page and their ranking on that page.

By leveraging the indexing power and ranking techniques of these engines Dogpile.com reaches more of the Web and is able to deliver the best results from across all these engines.

Finding: Dogpile.com covers the best of the best search results and returns a valuable mix of unique results deemed important by the top search engines.

- Results matched by 2 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.

The following chart outlines the results that Dogpile.com displays on its first result page.

Dogpile.com Total First Page Results for the 12,570 queries = 231,625

	% 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

Fig. 15

Dogpile.com Total First Page Non-Sponsored Results for the 12,570 queries = 145,529

	% 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

Fig. 16

Dogpile.com Total First Page Sponsored Results for the 12,570 queries = 40,786

	% 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

Fig. 17

The findings of this report highlight the fact that different search engines, which use different technology to find and present Web information, yield different first page search results. Metasearch technology brings all the information and views of different search engines together for the user's benefit. The fact that no one engine covers every page on the Internet, the majority of page one results are unique, and that almost half of the searches on the top four engines fail to elicit a click on a result offers a compelling case for using a metasearch engine that leverages the collective content and ranking methodologies of the major single source engines. Dogpile.com is in a unique position as the only search engine that sources Google, Yahoo!, MSN and Ask Jeeves. This exclusive offering results in the most comprehensive crawl of the Web and a search results set that highlights the best overall results from the Web.

Implications

There were three areas of implications gleaned from the results of this study. The implications centered on Web searchers, search engine marketers and metasearch technology.

Implications for Web Searchers

Either Search Multiple Search Engines or Use a Metasearch Engine like Dogpile.com
Web searchers are using an average of 2.8[10] search engines each month. This is highly inefficient for searchers quickly looking for the best results for their query.

There are many reasons for searching multiple search engines. This study did not set out to measure these. However, some examples of why people use multiple search engines may include:

- Couldn't find what was needed on one Web search engine
- Use of certain Web search engines for specific types of searches
- Just using the Web search engine that is most convenient at the time
- Desire to compare search results
- Aggregation of information around a specific topic
- Among others…

Many of the reasons for searching multiple search engines can be overcome. This study illustrates that using a metasearch engine that leverages the Web search power of the top Web search engines can reduce the time spent searching multiple search engines while providing then the best results from the top Web search engines in one place.

Implications for Search Engine Marketers

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, 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 potential missed audience by not leveraging the distribution power of both Google and Yahoo!. Those marketers who only optimize for, or purchase on, one search engine are missing valuable audience exposure by not running on both networks.

According to comScore Media Metrix 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[11]. Therefore, by only running ads on one of these engines, a marketer would miss out on millions of people 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.

Implications for Metasearch

Metasearch Engine's Leveraging Content from All the Top Engines Return Best Results of the Web
The results of this study highlight the fact that the top search engines (Google, Yahoo!, MSN and Ask Jeeves), have built and developed proprietary methods for indexing the Web and their ranking of keyword driven search results differs greatly.

Metasearch technology, especially from the industry leading metasearch engine Dogpile.com, harnesses the collective content, resources, and ranking capabilities of all four of the top search engines and delivers Web searchers a more comprehensive result set that brings the best results from the top engines to the first results page.

As indices continue to evolve, they will most likely remain differentiated. Since Web content is not static, there are barriers for any one engine's ability to cover the entire Web all of the time. As indices change and new Web content emerges, Dogpile.com's metasearch solution will be able to keep better pace with the Web as a whole than any single source search engine.

Conclusions

After 15 years of work, Web search is still in its infancy and technology around Web search will continue to evolve. The work done to date has uncovered four strong editorial voices for Web search based on unique ways of capturing and ranking search results. Google is different than Yahoo!. Yahoo! is different than Ask Jeeves. Ask Jeeves is different than MSN. These differences contradict the notion that all search engines are the same, and that searching one engine will yield the absolute best results of the Web. Through the relationships Dogpile.com has built with these engines, it is able blend these differences and provide the best results from the top Web search engines to the end user.

This study quantifies the similarities and, most importantly, the differences among the leading single source search engines. Each of the four single source engines measured has a unique voice and does a good job returning results they deem relevant based on that voice. The differences in indexing the Web and ranking results across these engines prevents users from feeling confident that they have found the best results for their search through the use of just one single source engine.

There are good search engines, but there is no perfect search engine on the Web. This is because intent is subjective to the end user, making it impossible for any search engine to understand every person's intent correctly each and every time. However, by using metasearch engine Dogpile.com, users can reduce the number of search engines they need to consult to one, making it easy to find the best results of the Web, and instilling confidence that they have performed the most comprehensive search of the Web with one click.

Resources

[1]comScore qSearch Data, May 2005, Custom Success Rate Analysis

[2,4,5,10] comScore Media Metrix, May 2005, US

[3]comScore Media Metrix, March 2005, World Wide

[6] Pew Internet & American Life Project: Search Engine Users, January 2005

[7]A. Gulli and A. Signorini. Building an open source metasearch engine. In 14th WWW, 2005.

[8]Search Engine Watch Newsletter, Chris Sherman, June 29, 2005

[9]comScore qSearch Data, May 2005

[10]InfoSpace internal log files, July 1-14 2005

[11] comScore qSearch data, Total US Internet Users - January 2005, InfoSpace analysis based on a custom report from comScore Marketing Solutions

Figure 1: Search Engine Watch Article, “Search Engine Sizes”, Danny Sullivan, January 28th, 2005.

Figure 2: A. Gulli and A. Signorini. Building an open source metasearch engine. In 14th WWW, 2005.

Figures 3 -12, 13-17: Dogpile.com Search Engine Overlap Study - July 2005, in collaboration with Dr. A. Spink (University of Pittsburgh) and Dr. J. Jansen (The Pennsylvania State University)

Figures 18-25: Screen shots taken of each search engine on June 15, 2005.

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

Ding, W., & Marchionini, G. (1996). A comparative study of Web search service performance. *Proceedings of the Annual Conference of the American Society for Information Science* (pp. 136-142).

Search Engine Watch Article, “Search Engine Sizes”, Danny Sullivan, January 28th, 2005.

Search Engine Watch Article, “How Search Engines Work, Danny Sullivan, October 14, 2002.

Appendix A

Control Analysis

The control analysis was conducted in April 2005 and was based off a sample of 10,316 user entered queries.

Great lengths were taken to ensure the methodology behind this study properly measured the overlap of search results. An exhaustive validation process was undertaken, prior to releasing the April 2005 study, which modeled various search result definitions to determine how to best measure overlap. Search result title, description, and display URL were all viewed as possible definitions of a search result. This study used the display URL for each search result.

There are known variations of display URLs used by some sites. A seemly unique display URL may link to the same page. Sites do this to better track where their traffic is coming from. Knowing this we set out to validate our overlap algorithm by applying various rules to the overlap algorithm.

To test our assumptions we applied the following rules to all first result page search results:

- Removed the domain prefix (www., www1., sports., search., etc)
- Removed the domain suffix and everything beyond (everything including the .com and beyond)

This resulted in only the root domain for the result. While this created false positive duplicates it completely mitigates any domain prefix variations sites may use which may otherwise we viewed as unique results.

Example: Results for keyword 'MLB'

> **Removing the domain prefix:**
> Result A: www. ebay.com/ would be considered the same as,
> Result B: www1. ebay.com/
>
> In addition, we truncated each display URL just after the .com. This mitigates any URL variation that may have over reported the number of unique results.

Example: Results for keyword 'MLB'

> **Removing the domain suffix:**
> Result C: Yahoo!.com/news would be considered the same as,
> Result D: Yahoo!.com/sports

Upon applying these rules and running the 10,316 queries, the data was found to be relatively unchanged. These rules, which by design would grossly over-estimate overlap, proved that the overlap definitions set forth in this study do in fact accurately measure search result overlap.

Finding: Overall overlap results remained relatively unchanged

- 3 engine overlap increased from 3% to 5%
- 2 engine overlap increased from 12% to 15%

Sample size: The sample of 10,316 keywords gives this study a +/- 1% margin of error.

Appendix B

Yahoo! Non-Sponsored Search Results

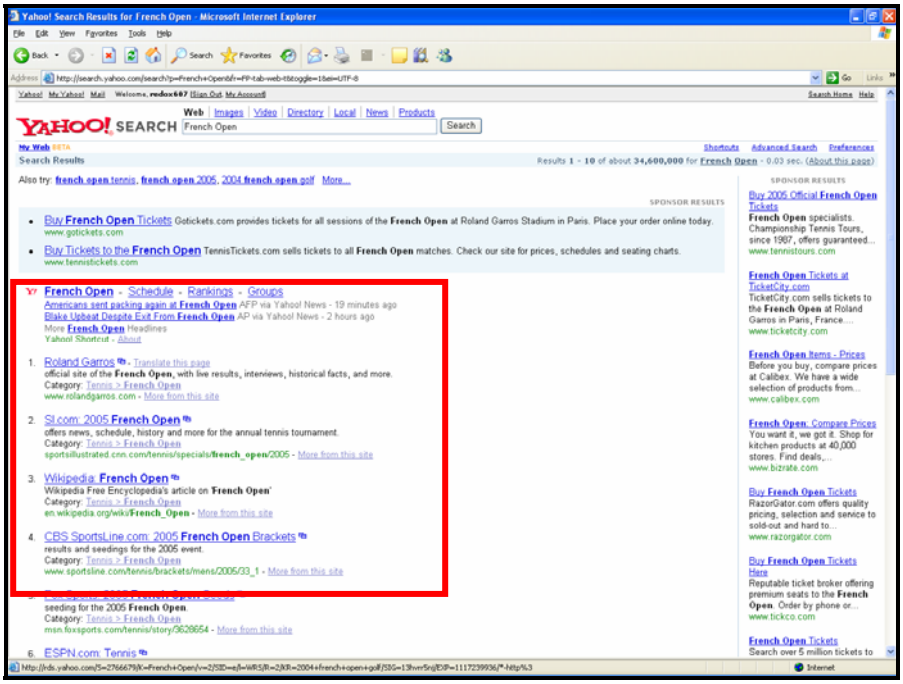

Fig. 18

Google Non-Sponsored Search Results

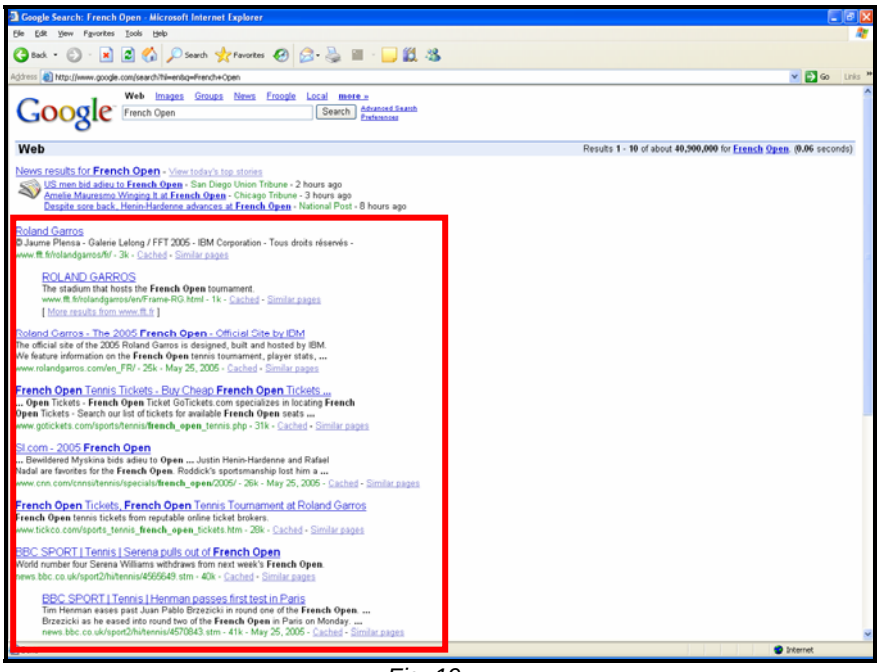

Fig. 19

MSN Search Non-Sponsored Search Results

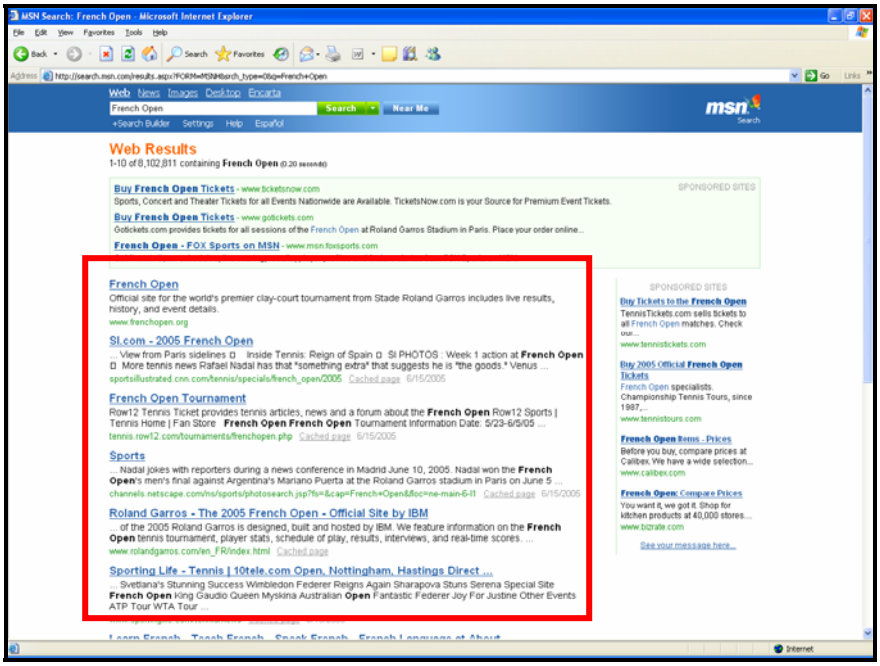

Fig. 20

Ask.com Non-Sponsored Search Results

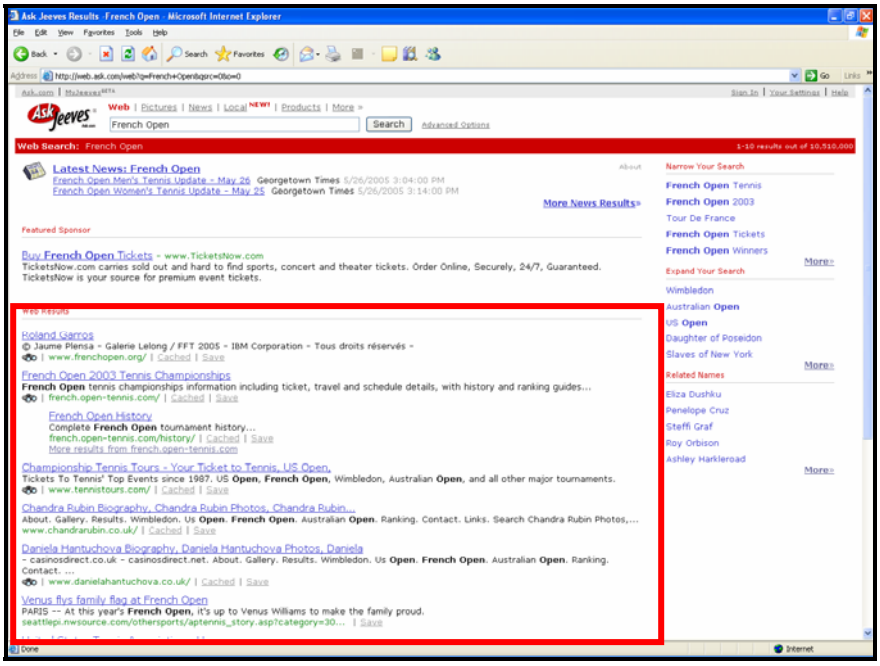

Fig.21

Yahoo! Sponsored Search Results

Fig. 22

Appendix C

Google Sponsored Search Results

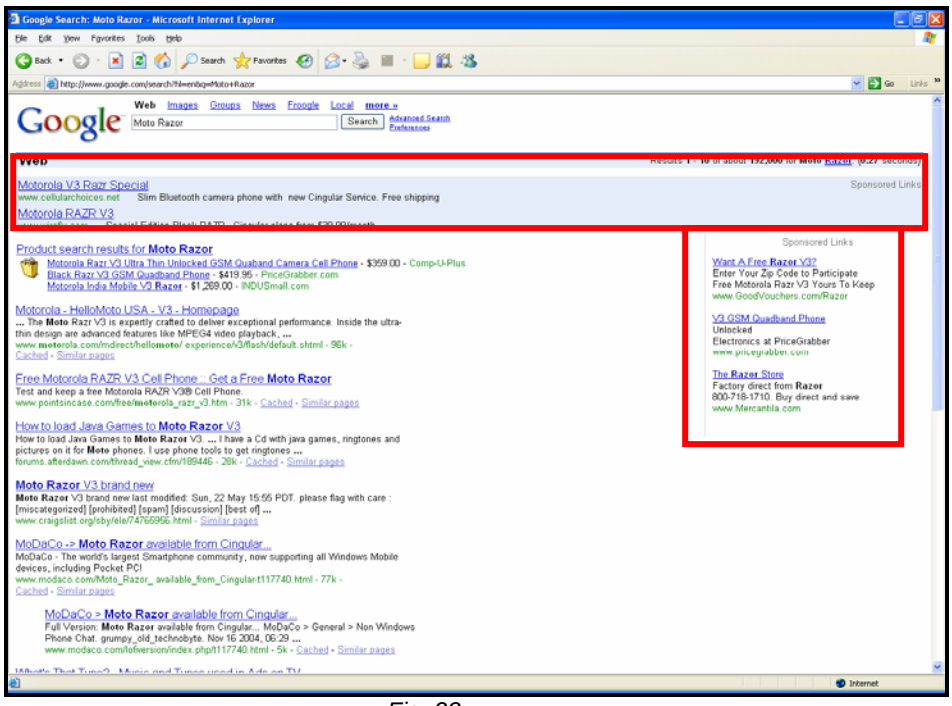

Fig. 23

Ask.com Sponsored Search Results

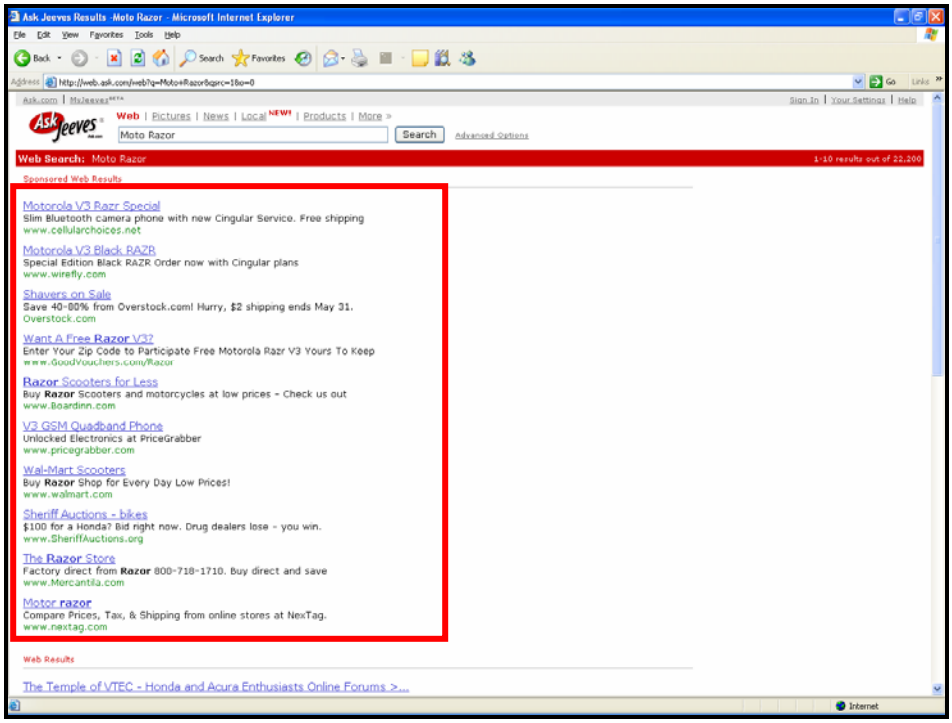

Fig. 24

MSN Search Sponsored Search Results

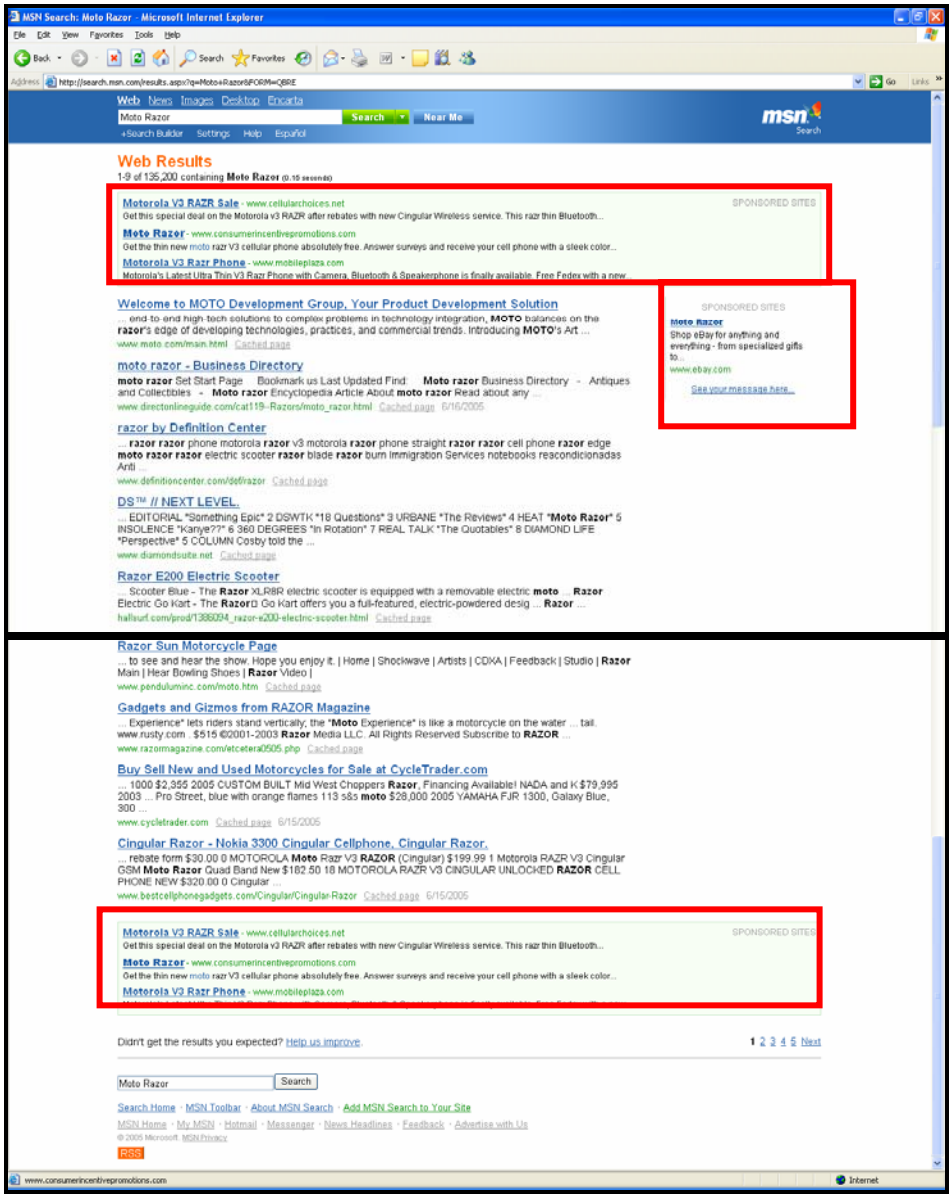

Fig. 25