Search Engines and Their Public Interfaces: Which APIs are the Most Synchronized?

Frank McCown
Old Dominion University
Computer Science Department
Norfolk, Virginia, USA 23529
fmccown@cs.odu.edu

Michael L. Nelson
Old Dominion University
Computer Science Department
Norfolk, Virginia, USA 23529
mln@cs.odu.edu

ABSTRACT

Researchers of commercial search engines often collect data using the application programming interface (API) or by "scraping" results from the web user interface (WUI), but anecdotal evidence suggests the interfaces produce different results. We provide the first in-depth quantitative analysis of the results produced by the Google, MSN and Yahoo API and WUI interfaces. After submitting a variety of queries to the interfaces for 5 months, we found significant discrepancies in several categories. Our findings suggest that the API indexes are *not* older, but they are *probably* smaller for Google and Yahoo. Researchers may use our findings to better understand the differences between the interfaces and choose the best API for their particular types of queries.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Experimentation, Measurement, Performance

Keywords

API, search engine

1. INTRODUCTION

In recent years, Google, MSN and Yahoo have developed freely available APIs for accessing their index, allowing researchers to more easily automate the data collection process and avoid breaking the stated policies of search engines that prohibit automated queries against their WUIs. Unfortunately, the APIs do not always give the same results as the WUIs. The listserves and newsgroups that cater to the API communities are full of questions regarding the perceived differences in results between the two, and only one limited study [4] we are aware of attempts to address the issue (only for the Google API).

Since none of the search engines publicly disclose the inner workings of their APIs, researchers are left wondering if the APIs are giving second-rate data. For example, Bar-Yossef and Gurevich [2] discount their findings, stating their belief that the APIs are "served from older and smaller indices than the indices used to serve human users." Other researchers appear to believe that the APIs serve the same data as the WUIs [5]. The purpose of this study was to test these assumptions.

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

2. EXPERIMENT

Every day for five months (late May to Oct 2006) we submitted four types of queries (3500 total) to each interface:

1. General search terms. We queried for the top 100 results and the total number of results using 50 popular search terms[1] and 50 computer science (CS) terms[2].

2. URL backlinks. We queried for the number of backlinks to 100 randomly selected URLs.

3. Pages indexed for a website. We asked how many pages were indexed for 100 randomly selected websites.

4. URL indexing and caching. We queried to see if 100 randomly selected URLs were indexed and/or cached.

We used three distance measures to compare the top search results for our popular and CS terms: overlap P, Kendall tau for top k lists K [3], and measure M [1]. Each measure was normalized so 1 meant complete agreement and 0 complete disagreement. P only penalizes for non-shared results, but K additionally penalizes for results that are out of position. M penalizes changes in the top of the list more heavily than changes at the bottom (since humans typically examine the top results more often than the bottom ones). We refer the reader to the cited works for formal definitions.

We initially compared WUI results from day n to day $n-1$ and API results from day n to day $n-1$ and found a great amount of fluctuation. The interfaces tended to change at the same rate each day for most terms, but we found several popular term results like *carmen electra* and *jessica simpson* that appeared to change at very different rates in both of Google's interfaces. We did not observe this phenomena in Yahoo and MSN. The K means of both interfaces for Google, MSN and Yahoo were 0.94, 0.94 and 0.95, respectively.

Comparing the WUI and API results each day, we see a great deal of variation (Figure 1)[3]. M appears significantly lower than the other measures when examining Google's results for both popular and CS terms and for MSN popular terms, suggesting that Google's and MSN's top results are the most significantly different between their interfaces. Google's distance measures are all significantly higher for

[1] `http://50.lycos.com/`
[2] `http://en.wikipedia.org/wiki/List_of_cs_topics`
[3] The MSN gap was due to MSN invalidating our API key by mistake for 17 days.

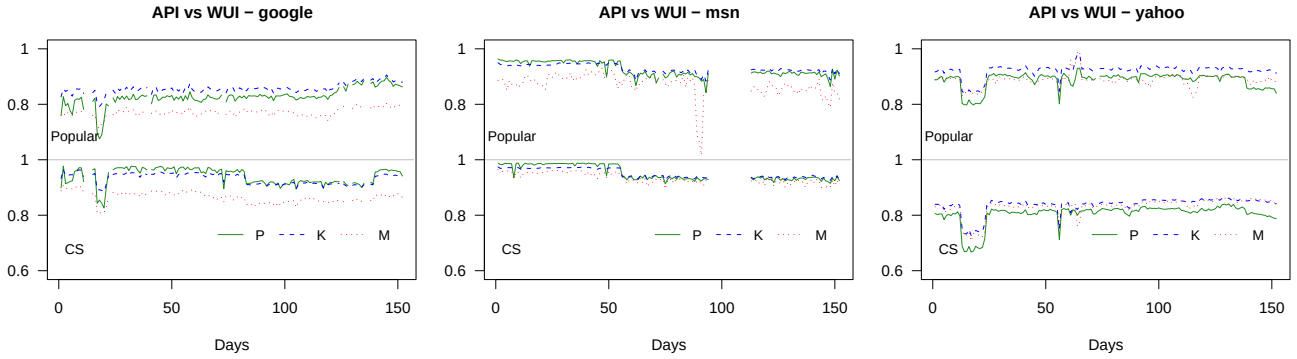

Figure 1: Distance between WUI and API top 100 search results.

Table 1: Loose Disagreements (Means)

		Total results	Total backlinks	Pages indexed
Google	API > WUI	7.9%	0.6%	4.9%
	WUI > API	46.5%	1.5%	46.0%
MSN	API > WUI	0.9%	2.2%	5.4%
	WUI > API	0.6%	21.4%	7.3%
Yahoo	API > WUI	1.0%	24.8%	14.1%
	WUI > API	37.5%	28.1%	8.6%

CS terms than popular terms, but the reverse is true for Yahoo (two-sample Wilcoxon signed rank test, $p < 0.001$). These differences may be explained by how Google and Yahoo treat web spam. Averaging the distance measures together, Google's and Yahoo's API results are 14% different than the WUI results, and MSN is 7% different.

We tested to see if the API index was older (or newer) than the WUI index by comparing API results from day n to WUI results on day $n \pm 1$, $n \pm 2$, etc. For all three search engines, we found the highest correlation on day n. Therefore the API indexes are not older or newer.

We examined the number of times the interfaces produced top k results that were identical in rank ($K = 1$) or in set membership ($P = 1$). Google never produced identically ranked top 100 results, and MSN and Yahoo did only slightly better (0.2%). For top 10 results, MSN (38%) and Yahoo (32%) improve considerably, but Google only slightly (4%). There is only modest improvement of 3-6% by all of the search engines when we disregard ranking.

When we examined the total results, backlinks and pages indexed, we rarely saw exact agreement, so we examined *loose disagreements*, when the API value is greater than or less than ± 10% of the WUI value. We summarize our findings in Table 1. Responses to the URL indexing and caching queries resulted in far more consistent interface responses for all search engines; Google disagreed only 1.0% of the time, MSN 1.1% and Yahoo 6.8%.

3. CONCLUSIONS

Researchers should expect that the results obtained through any search engine's API will rarely be identical to what the general WUI user sees. This especially impacts applications where recall is important. But some APIs perform better in certain categories as we have summarized in Table 2. MSN appears to have the most synchronized interfaces overall.

Table 2: Synchronized Interfaces

Type	Most synched	Least synched
Search for popular terms	MSN	Google
Search for CS terms	MSN	Yahoo
Total results	MSN	Google
Total backlinks	Google	Yahoo
Pages indexed per website	MSN	Google
Indexed/cached status	Google/MSN	Yahoo

We have shown that all three search engines provide nearly synchronized changes in their results each day, and the API results are most similar to the WUI results on the same day. Therefore we conclude that the APIs are *not* serving from an older index. It is possible though that Google and Yahoo are serving from a *smaller* index; both WUIs consistently report total results that are higher than the APIs (of course we cannot directly verify these estimates, and Google adds 'supplemental results' to their WUI results). Yahoo's WUI consistently reports larger backlink counts, and Google's WUI consistently reports larger website page counts. To give a more definitive answer, we suggest a future experiment that randomly samples from each corpus and compares the overlap [2].

On December 20, 2006, Google announced they were beginning to deprecate their SOAP API in favor of an AJAX API, a move which may require researchers to readopt old screen-scraping methods. Despite this setback, it is our hope that commercial search engines will make a committed effort to provide more synchronized interfaces for the academic community in the future.

4. REFERENCES

[1] J. Bar-Ilan, M. Mat-Hassan, and M. Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, July 2006.

[2] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *Proceedings of WWW '06*, pages 367–376, 2006.

[3] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[4] P. Mayr and F. Tosques. Google Web APIs - an instrument for webometric analyses? In *Proceedings of ISSI 2005*, 2005.

[5] M. Thelwall. Can the Web give useful information about commercial uses of scientific research? *Online Information Review*, 28:120–130, 2004.