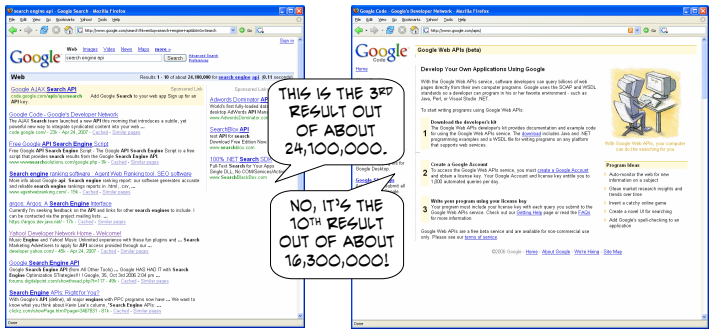
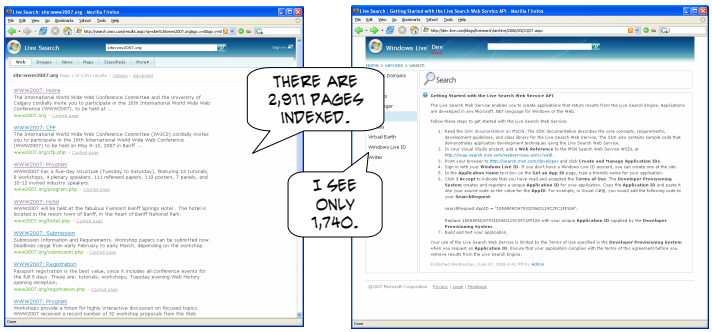
Search Engines and their Public Interfaces: Which APIs are the Most Synchronized?

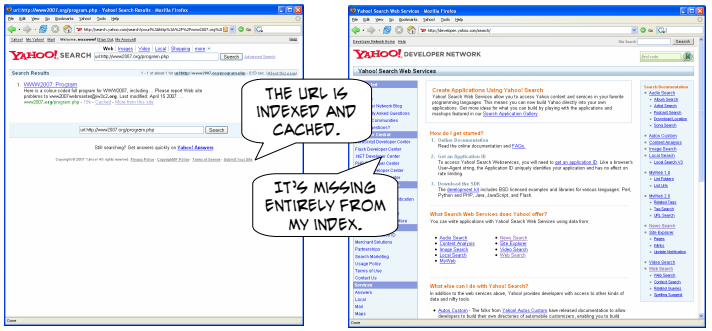
Frank McCown and Michael L. Nelson
Department of Computer Science, Old Dominion University, Norfolk, Virginia, United States

Researchers: Screen Scrape or Use the APIs?

Web User Interface (WUI)	Application Programming Int. (API)

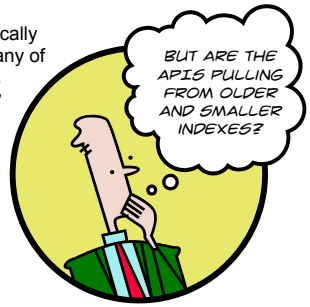






Google **Terms of Service:** "You specifically agree not to access (or attempt to access) any of the Services through any automated means (including use of scripts or web crawlers…)"

Windows Live **Terms of Service:** "You may not… use any automated process or service to access and/or use the service (such as a BOT, a spider, …)"



5 Month Experiment

Late May to Oct 2006:

1. **General search terms**. Queried for the top 100 results and total results using 50 popular search terms and 50 computer science (CS) terms.
2. **URL backlinks.** Queried for the number of backlinks (inlinks) to 100 randomly selected URLs.
3. **Pages indexed for a website.** Asked how many pages were indexed for 100 randomly selected websites.
4. **URL indexing and caching.** Queried to see if 100 randomly selected URLs were indexed and/or cached.

Comparing Search Results

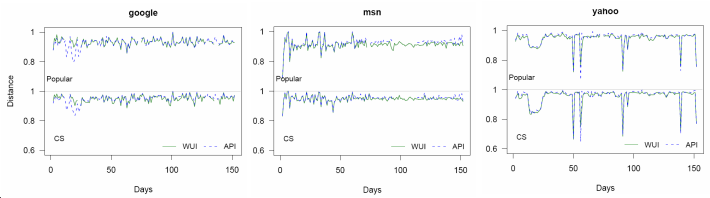
1. Overlap (P)
2. Kendall tau for top *k* results (K)[1]
3. Penalize changes at the top more heavily (M)[2]

0 → 1 More similar

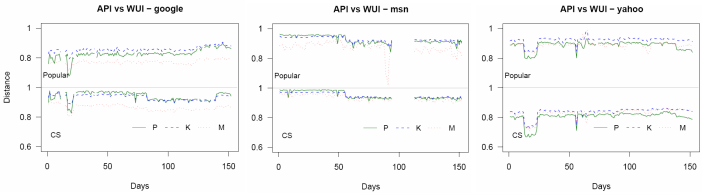
[1] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top *k* lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.
[2] J. Bar-Ilan, M. Mat-Hassan, and M. Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, July 2006.

1. ABCD 2. EDAF	*Examples*	1. ABCD 2. AEDF
$P = 0.50$ $K = 0.44$ $M = 0.14$		$P = 0.50$ $K = 0.56$ $M = 0.66$

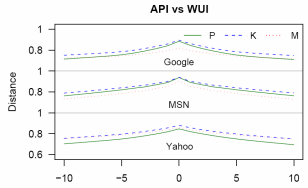
Comparing WUI to WUI & API to API on Successive Days



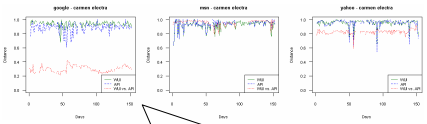
COMPARING WUI TO API



FOR ALL 3 SEARCH ENGINES, THE WUI & API ARE MOST SYNCHRONIZED ON THE **SAME DAY**.

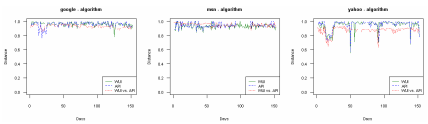


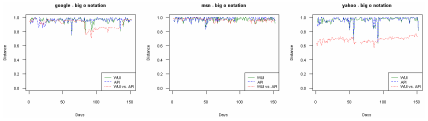
EXAMPLES



MSN IS MOSTLY SYNCHRONIZED FOR "ALGORITHM".

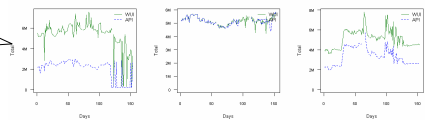
GOOGLE IS LESS SYNCHRONIZED FOR POPULAR TERMS.

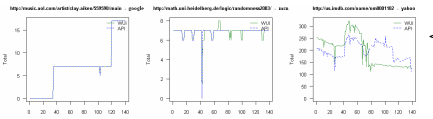




YAHOO IS LESS SYNCHRONIZED FOR CS TERMS.

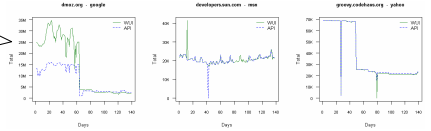
HOW MANY TOTAL RESULTS ARE THERE?





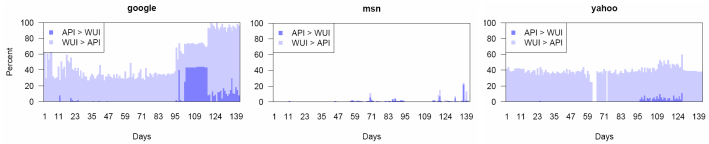
WHOSE BACKLINK COUNTS ARE CORRECT?

GOOGLE'S API SHOWS FEWER PAGES INDEXED.

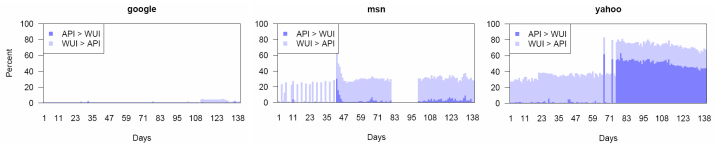


LOOSE DISAGREEMENTS

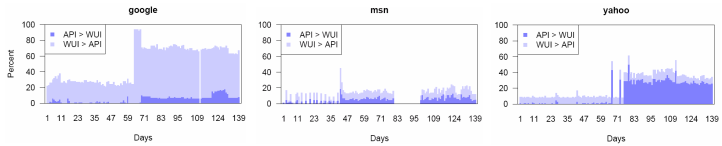
TOTAL SEARCH RESULTS



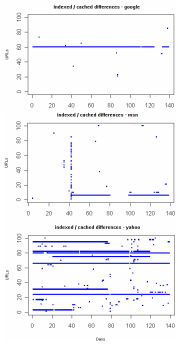
TOTAL BACKLINKS



TOTAL INDEXED PAGES PER WEBSITE



INDEXED / CACHED DISAGREEMENTS



GOOGLE & YAHOO **MIGHT** BE PULLING FROM **SMALLER** INDEXES.

YAHOO SEEMS TO BE CONFUSED.

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%

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

SEE ALSO

Frank McCown and Michael L. Nelson. Agreeing to Disagree: Search Engines and their Public Interfaces. ACM IEEE Joint Conference on Digital Libraries (JCDL 2007). *To appear.*

All graphs: http://www.cs.odu.edu/~fmccown/research/se_apis/
Complete data set available upon request.

Other research projects at Old Dominion University:

mod_oai: An Apache Module for Efficient, Automatic Web Harvesting
http://www.modoai.org/

Lazy Preservation: Reconstructing Websites from the Web Infrastructure
http://www.cs.odu.edu/~fmccown/warrick/

WARRICK

OldDominionUNIVERSITY
Changing Lives