RESEARCHERS: SCREEN SCRAPE OR USE THE APIs?

WEB USER INTERFACE (WUI)  APPLICATION PROGRAMMING INT. (API)

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)
3. Penalize changes at the top more heavily (M)

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...)

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.

Table 1: Loose Disagreements (Means)

Table 2: Synchronised Interfaces

SEE ALSO


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

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/