Agreeing to Disagree: Search Engines and Their Public Interfaces

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

Google, Yahoo and MSN all provide both web user interfaces (WUIs) and application programming interfaces (APIs) to their collections. Whether building collections of resources or studying the search engines themselves, the search engines request that researchers use their APIs and not "scrape" the WUIs. However, 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. We have queried both interfaces for five months and found significant discrepancies between the interfaces in several categories. In general, we found MSN to produce the most consistent results between their two interfaces. Our findings suggest that the API indexes are *not* older, but they are *probably* smaller for Google and Yahoo. We also examined how search results decay over time and built predictive models based on the observed decay rates. Based on our findings, it can take over a year for half of the top 10 results to a popular query to be replaced in Google and Yahoo; for MSN it may take only 2-3 months.

Categories and Subject Descriptors: H.3.5 [Information Storage and Retrieval] Online Information Services: Web-based services

General Terms: Experimentation, Measurement

Keywords: API, distance measurement, search engine results, search engine interfaces

1. INTRODUCTION

Commercial search engines have long been used in academic studies. Sometimes the search engines themselves are being studied, and sometimes they are used to study the Web or web phenomena. In the past, researchers have either manually submitted queries to the search engine's web user interface (WUI), or they have created programs that automate the task of submitting queries. The returned re-

sults have been processed manually or by programs that rely on brittle screen-scraping techniques.

But data collection mechanisms for search engines have changed in the past few years. Three of the most popular commercial search engines (Google, MSN and Yahoo) have developed freely available APIs for accessing their index, and researchers can now use these APIs in their automated data collection processes. Unfortunately, the APIs do not always give the same results as the WUI. The listserves and newsgroups that cater to the API communities are full of questions regarding the perceived differences in results between the two. None of the search engines publicly disclose the inner workings of their APIs, so users are left wondering if the APIs are giving second-rate data. This anecdotal evidence has led some researchers to question the validity of their findings. For example, Bar-Yossef and Gurevich [8] state that "due to legal limitations on automatic queries, we used the Google, MSN, and Yahoo! web search APIs, which are, reportedly, served from older and smaller indices than the indices used to serve human users." Other researchers may be totally unaware of the differences. When writing about a 2004 experiment using the Google search engine, Thelwall [47] stated that "the Google API... could have been used to automate the initial collection of the results pages, which should have given the same outcome" as using the WUI.

The main purpose of this study is to examine the differences between what is reported between the WUI and API when queried with a variety of queries that researchers and users frequently use. Our findings allow us to address the question of whether the APIs are pulling from older and smaller indexes. A secondary purpose is to examine how search results decay over time and provide predictive models for determining the half-lives of search results.

2. BACKGROUND AND RELATED WORK

2.1 Search Engine APIs

In early 2002, Google generated a great deal of fanfare when it became the first search engine to release a free SOAP-based API for accessing its index. Google required users to register for a license key that allowed 1000 requests to be issued per day. Yahoo and MSN released public APIs in early and late 2005, respectively, but in an effort to out-do Google, they greatly reduced the usage restrictions of their APIs. Yahoo's REST-based API allows 5000 daily requests per IP address, per application ID [51]. MSN's SOAP API allows 10,000 daily requests per IP address, per application

ID [40]. By allowing more requests per day and only enforcing the quota to a particular IP address, Yahoo and MSN allow developers to create applications for third parties which do not require users to register for license keys.

All three search engines have expanded their APIs to include facilities like maps, spelling, news, etc. Despite their popularity (or because of it), none of the search engines give direct technical support for their APIs; there are newsgroups and listserves that developers can use, and representatives from the search engines keep an eye on the discussions and occasionally respond.

Researchers have been attracted to using the APIs not only because it eases data collection activities, but because of the legal restrictions against automating data collection using the WUIs of several search engines. According to Google's terms of service [23], users "may not send automated queries of any sort to Google's system without express permission in advance from Google." Ask.com and MSN have similar policies [1, 39].

2.2 Research Use of the APIs

Prior to the release of the search engine APIs, collecting distributed web content to build a digital library (DL) required "focused crawling" – crawling the Web for content matching the focus of a particular DL [9, 10]. While many DLs still use this technique, focused crawling is now often augmented [25, 42, 52], or even replaced [27, 30], with search engine APIs. Depending on their collection development policy, DL administrators should be informed about the issues of completeness and synchronization between the API and WUI interfaces.

There are numerous studies which have evaluated search engines exclusively through either the WUI or API. For example, early studies on the size of the indexable web and search engine overlap automated thousands of queries against the WUI when no API was available [11, 31]. A 2005 study of search engine size and overlap [24] also used the WUI exclusively, although recent measurements taken in [8] used only the APIs. We found numerous studies in the literature that have used the Google API to perform general search queries where ranking of the items is important [14, 26, 29, 35, 44], search for backlinks [41] and determine if particular content was indexed or not [8, 37, 46]. In most of these studies, the findings would be altered if the Google API reported significantly different results than the WUI.

Evaluation of API search results has received little attention in the literature. Bar-Ilan [5] provided an excellent summary of current search engine deficiencies in 2005, but she did not evaluate any search engine APIs. Mayr and Tosques [33] performed a limited study of the Google API in a 2004-2005 experiment. They found a large discrepancy between the results obtained between the Google API and WUI; the WUI produced as much as six times the number of hits for the word *webometrics* over the period of a year. They also found more differences in the search results obtained from the API than the WUI, but they do not provide any measurements.

A number of studies compare the results obtained from one search engine with others when given the same query [4, 6, 18, 19, 45], or they examine how the results change over time [2, 3, 6, 7, 48]. The studies may use a handful of queries (e.g., [4]) or a very large data set (e.g., [45]), and typically just the top 10 to 50 results are examined since these are the results most frequently examined by human searchers. While a few studies employ humans to judge the relevance or ranking of the results (e.g., [48]), most compare the results by applying a number of difference measures like overlap, Kendall tau, Spearman's footrule, and other measures that take into account rank ordering. We have employed these distance measures in our study and do not attempt to examine relevance or technical precision [3]. All these studies have concluded that most search engines produce very different results when given the same query (prompting active research into meta-searching [21, 45]) and that the results also change at different rates over time. All of the studies cited have used results obtained using the WUI interfaces, and none of them have examined how the results would be different if using the search engine APIs. We are unaware of any studies that have used queries to measure website indexing or backlinks over time.

2.3 Comparing Search Engine Results

Most of our analysis focuses on comparing the top 10 and 100 search results obtained using a variety of query terms. In order to compare the results over time and the differences between the two interfaces, we review three distance measures which have been used in similar studies to compare top k search results. All three measures range from 0 to 1 where 0 means complete disagreement between the results and 1 means complete agreement.

As mentioned in Section 2.2, computing overlap is a common way to compare search results. The formal definition we will use for the normalized overlap P of two top k lists (lists of size k) τ_1 and τ_2 is: $P = |\tau_1 \cap \tau_2|/k$.

For example, consider the following two lists of top 4 search results where each letter represents a unique result:

1. ABCD
2. EDAF

In this case result A was returned first in list 1 and third in list 2, B was returned second in list 1 but is not present in list 2, etc. The overlap P for these two lists is $2/4 = 0.5$.

Overlap does not convey changes in ranking but merely set membership. A well-known measure like Kendall tau distance is a stronger distance measure that takes into account changes of rank between two lists. Kendall tau distance counts the number of pairwise disagreements (discordant pairs) between two lists and is equivalent to the number of swaps needed to transform one list to another using the bubblesort algorithm. Kendall tau distance K_τ for two lists τ_1 and τ_2 is defined:

$$K_\tau(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2) \qquad (1)$$

where P is the set of unordered pairs of distinct elements in τ_1 and τ_2, and $\bar{K}_{i,j}(\tau_1, \tau_2) = 1$ if i and j are in the opposite order in τ_1 and τ_2 and 0 if they are in the same order.

As in our example, there are often results that are not shared between two lists of search results. Since Kendall tau distance assumes both lists have all elements in common, we must modify it for use on top k lists. Fagin et al. [19] have developed an equivalence class of distance measurements based on Kendall's tau and Spearman's footrule that can be applied to top k lists that do not necessarily share all

elements. Fagin et al. show that all the distance measures they developed are essentially the same, and therefore it is not important which one is used. We use the Kendall tau distance $K^{(p)}$ with penalty $p = 0$, what Fagin et al. call the "optimistic approach".

Assume P is the set of all unordered pairs of distinct elements in $\tau_1 \cup \tau_2$ where τ_1 and τ_2 are two top k lists. $K^{(0)}$ is defined formally as

$$K^{(0)}(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2) \qquad (2)$$

where

- $\bar{K}_{i,j}(\tau_1, \tau_2) = 1$ if any of the following conditions hold:
  1. i and j are in both lists and in the opposite order
  2. i and j are in the same list with j ranked ahead of i, and only i is in the other list
  3. i appears only in one list and j is only in the other
- $\bar{K}_{i,j}(\tau_1, \tau_2) = 0$ for all other conditions

In order to compare $K^{(0)}$ with our overlap distance measure, we need to normalize $K^{(0)}$ so two identical lists have a value of 1, and two lists that share no elements have a value of 0. If two top k lists have no shared elements, then $K^{(0)}$ would be equal to k^2. Therefore our normalized version of $K^{(0)}$, which we denote K, becomes:

$$K = 1 - \frac{K^{(0)}(\tau_1, \tau_2)}{k^2} \qquad (3)$$

To illustrate how K can be intuitively computed using K_τ, consider again the top 4 search results from the previous example. Since B and C do not appear in list 2, we assume optimistically that they have slid down to positions 5 and 6 in list 2, just beyond the bounds of our top 4 list. Since E and F do not appear in list 1, we again assume that they are at positions 5 and 6 in list 1. So to compare these lists with K_τ, we take the unshared elements of list 1 and append them to list 2 and vice versa. The two lists become:

1. ABCDEF
2. EDAFBC

We can now compute $K_\tau(\tau_1', \tau_2')$ on the altered lists since they now have all elements in common. In this case there are 9 pairwise disagreements. Since $K_\tau(\tau_1', \tau_2')$ and $K^{(0)}(\tau_1, \tau_2)$ are equivalent, K is $1 - 9/16 = 0.4375$.

The third distance measure we will use in our experiment is the M measure developed by Bar-Ilan et al. [7]. The M measure is based on findings of an eye-tracking study [17] that showed participants where much more likely to view the top set of search results on a page than the bottom. The M measure encodes the intuition that changes in the top positions of the search results should be weighed more heavily than changes at the bottom of the result set. The unnormalized value M' for two top k lists τ_1 and τ_2 is defined:

$$M'(\tau_1, \tau_2) = \sum_{\tau_1 \cap \tau_2} \left| \frac{1}{rank_1(i)} - \frac{1}{rank_2(i)} \right|$$
$$+ \sum_{\tau_1 \setminus \tau_2} \left(\frac{1}{rank_1(i)} - \frac{1}{k+1} \right) \qquad (4)$$
$$+ \sum_{\tau_2 \setminus \tau_1} \left(\frac{1}{rank_2(i)} - \frac{1}{k+1} \right)$$

where $rank_j(i)$ is the rank of element i in τ_j (e.g., $rank_1(\text{D})$ is 4 from our previous example).

To normalize M' so the value is 1 when the two lists are identical and 0 when the two lists share no common elements, we compute the normalization factor n to be 4.03975 for $k = 10$ and 8.39456 for $k = 100$:

$$M = 1 - \frac{M'(\tau_1, \tau_2)}{n} \qquad (5)$$

Returning again to our previous example, M would be $1 - 2.2/2.56667 = 0.1429$ where $n = 2.56667$ is the value of M' when two top 4 lists share no elements. Here we see M assigning a much lower distance to the two lists than did P (0.5) or K (0.4375) because of large changes to the first couple of elements. If A had remained the first element in list 2 (AEDF), P would remain 0.5, but K would now be 0.5625, and M would be 0.6623.

3. EXPERIMENT SETUP

We devised an experiment to compare the WUI and API results based on four types of queries:

1. **General search terms** - query for the top 100 results and the total number of results
2. **URL backlinks** - query for the number of backlinks to a particular URL
3. **Total URLs indexed for a website** - query for the number of pages indexed for a website
4. **URL indexing and caching** - query if a particular URL is indexed and/or cached

We obtained 50 popular search terms from Lycos Top 50 [32] and 50 computer science (CS) terms obtained from a list at Wikipedia [50]. We chose two different types of terms in order to see if the term type had any affect on the stability of the search results and the synchronization of the search engine interfaces. For each search term we obtained the first 100 results produced by each interface. We ignored all sponsored results from the WUIs (APIs do not give sponsored results). We did not use any advanced syntax in our searches. Care was taken to ensure that both interfaces were asked to return the same results (no filtering, similar results or restrictions). Although Google often returns 'supplemental results' [49] in their WUI results but not in their API results, there is no way to control for it.

At the outset of our experiment, we randomly selected 100 URLs from the top 100 results generated by all the search terms. We used these URLs to ask 1) if the given URL was indexed by the search engine (using 'info:' for Google and 'url:' for MSN and Yahoo), and 2) if there was a cached URL for the given URL. We also asked each search engine to report the number of backlinks it had recorded for each of these URLs using the 'link:' query. We used the website domain name of the same 100 randomly obtained URLs to ask each search engine the number of web pages it had indexed for the website. For example, if the URL was `http://foo.org/abc.html`, we would query for 'site:foo.org'. The same URLs were used throughout the experiment.

We used a single server (beatitude.cs.odu.edu with IP address 128.82.4.22) to automate our data collection activities. Each of our queries were issued serially, first to the WUI and then to the API. The queries were executed beginning at 2

Table 1: Statistics for Top 100 Search Results Comparing WUI to WUI and API to API

Type	Dist	Interface	Google			MSN			Yahoo		
			Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Popular	P	WUI	0.92	0.49	1	0.89	0	1	0.93	0.31	1
		API	0.91	0.30	1	0.90	0.19	1	0.94	0.32	1
	K	WUI	0.92	0.53	1	0.91	0.28	1	0.95	0.44	1
		API	0.93	0.39	1	0.93	0.28	1	0.95	0.35	1
	M	WUI	0.94	0.45	1	0.84	0.16	1	0.95	0.34	1
		API	0.93	0.22	1	0.89	0.05	1	0.95	0.34	1
CS	P	WUI	0.94	0.53	1	0.93	0.24	1	0.94	0.29	1
		API	0.93	0.47	1	0.93	0.24	1	0.95	0.31	1
	K	WUI	0.95	0.62	1	0.95	0.31	1	0.95	0.36	1
		API	0.95	0.59	1	0.95	0.32	1	0.96	0.39	1
	M	WUI	0.94	0.34	1	0.93	0.26	1	0.95	0.35	1
		API	0.95	0.43	1	0.94	0.05	1	0.95	0.35	1

am in the morning (Eastern Standard Time) when search engine traffic is typically light [43]. We delayed a random 15-60 seconds per query to each WUI to avoid generating too much traffic in a small amount of time and being detected as an automated script. Google and Yahoo have both taken steps in the past few years to deny access to IP addresses where they have detected what they believe to be automated queries [22, 38]. These efforts are likely in response to aggressive querying from the SEO community and from viruses that use search engines to find new targets [20].

We issued a total of 3500 queries each day to the three search engines. The WUI queries were aimed at the US version of each search engine: www.google.com, search.msn.com and search.yahoo.com. We did not target specific datacenters since this is problematic and atypical of general users' behavior [16]. All WUI web page responses were archived in case changes to the HTML format broke any of our screen-scraping regular expressions (this happened a couple of times with MSN and Yahoo). The WUI queries produced roughly 50 MB of data (5.5 MB compressed) daily. We issued queries from late May 2006 through Oct. 2006.

4. EXPERIMENT RESULTS

4.1 Query Errors

We encountered numerous transitory errors throughout the experiment. The Yahoo API typically generated a few "Bad Request: service temporarily unavailable" responses each morning. On occasion the error would be received 20-30 times on the same day. Towards the end of the experiment, Google's API frequently returned 502 "Bad Gateway" errors. These errors became so prevalent that we modified our scripts to delay for 15 seconds and re-submitted our query each time the 502 error was returned. On rare occasions, the WUI of all three search engines would respond with an http 500 response.

Throughout our experiment, the Google API frequently returned back an error response ("Exception from service object: For input string XYZ") for the queries *list* and *database*. This error was caused on Google's back end when they attempted to return a total result value like 3,450,000,000 that was too large for the 32-bit integer used by their API.

The most troubling error was incurred on Aug. 29 when MSN invalidated our license key without warning. For a period of 17 days we received an "Invalid value for AppID in request" error for every API query before replacing the

key with a new one. We received this error several more times throughout the experiment, but our key appeared to function properly for the vast majority of queries.

4.2 Search Term Results

4.2.1 Intra-Interface Comparisons

When examining the top k search results produced each day from popular and CS terms, we used the three distance measures discussed in Section 2.3: overlap P, Kendall tau for top k lists K and measure M. We applied these measures using intra-interface comparisons (WUI vs. WUI results and API vs. API results) and comparisons between interfaces.

First we examine how the top 100 WUI and API results change over time. We averaged the distances for all search results, separating the CS and popular term results. Table 1 shows the descriptive statistics when comparing the WUI results on day n to day $n-1$ and the API results on day n to day $n-1$. In Figure 1 we plot the distance measures for Google, MSN and Yahoo using the K distance (the API gap for MSN was due to the invalid key error discussed in Section 4.1). Graphs using the other distance measures looked very similar in terms of how the WUI and API distance lines closely followed each other. Excluding days 15-25 for Google, the distance lines for all three search engines appear to move in synchronization, even when very large changes take place. For example, we see a spike on day 50 for Yahoo where the WUI and API results both report less than a 0.7 K distance with day 49 results. We see very similar results when examining the top 10 results, and we refer the reader to [34] for detailed findings.

For the most part, the CS and popular term results averaged the same degree of change for both Google and Yahoo. Only MSN reported higher average distance measures for CS results. The averages reported by both interfaces were largely the same; MSN reported the greatest average disagreement between the interfaces using the M measure for popular term results, a difference of 0.05.

When examining each term's results individually, we found that several of the popular term results did not have synchronized interface updates in Google. Google's API results for *carmen electra* and *jessica simpson*, for example, appeared to change independently from the WUI results. All 50 of the CS terms showed close change synchronization between the Google WUI and API results as did all the Yahoo and MSN results for all terms.
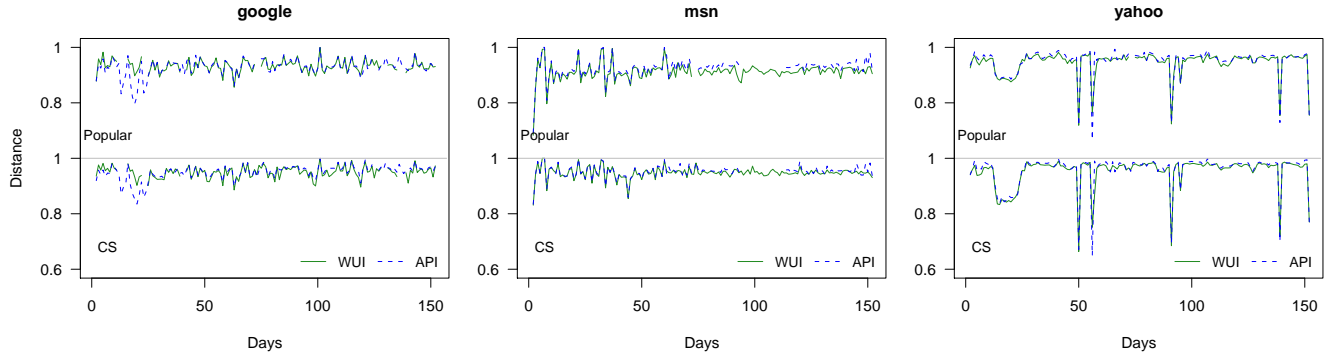
Figure 1: K **distance between top 100 search results when comparing day n to day $n-1$.**

4.2.2 WUI vs. API Comparisons

We have seen that the daily results produced by the interfaces changed at nearly the same rate for all three search engines. Now we examine how similar the WUI and API results are to each other on any given day. Figure 2 (next page) plots the distance between the API and WUI results on each day using all three distance measures. Descriptive statistics are given in Table 2. When we examined just the top 10 results, we found the measurements to be nearly the same (less than a change of 0.05) except for Google whose popular and CS P averages dropped by 0.13 and 0.15, respectively, and whose K average dropped by 0.6.

The M measure appears significantly lower than the other measures when examining Google's results for both popular and CS terms (Table 2). MSN also appears to have lower M measures for popular terms. This suggests that the top results are the most significantly different between the interfaces for both Google and MSN. Averaging the distance measures together for top 100 results, Google's and Yahoo's API results are 14% different than their WUI results, and MSN's is 7% different. For top 10 results, Google's API is 20% different, Yahoo's is 14%, and MSN's is 8%.

We also see from Table 2 that Google's overall distance measures are higher for CS terms than popular terms, but the reverse is true for Yahoo. A two-sample Wilcoxon signed rank test confirms the significance at the $p < 0.001$ level. Although the mean differences between the term type results are also significantly different for MSN ($p < 0.001$), the values are much closer.

In an effort to see if the API results were older (or newer) than the WUI results for any given day, we compared each set of top 100 results on day n with day $n\pm1$, $n\pm2$, etc. The results are graphed in Figure 3. For all three search engines, the highest degree of similarity occurs on the same day (offset 0) which means that the API results are *not* pulling from an older (or newer) index. This is true when examining the top 10 results as well.

The WUIs and APIs rarely produced identical top 100 results. In Table 3 we show the percentage of times that the interfaces produced top k results that were either identical in rank ($K = 1$) or identical in set membership ($P = 1$). Not once in 5 months did the Google API ever produce a single top 100 result that was identical to the Google WUI. MSN and Yahoo did only slightly better (0.2%). When we exam-

Table 2: Statistics for Top 100 Search Results Comparing WUI to API

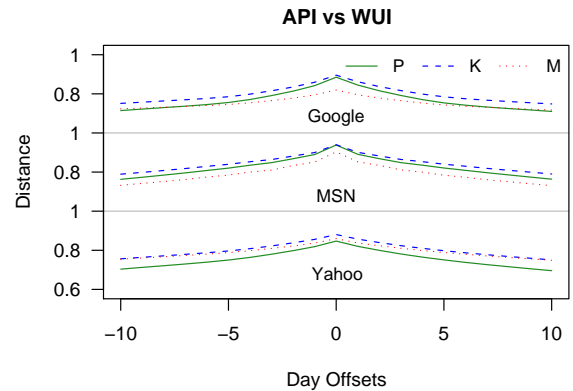| Type	SE	Dist	Mean	Min	Max
Popular	Google	P	0.83	0.05	1
		K	0.86	0.07	0.99
		M	0.77	0.01	0.99
	MSN	P	0.93	0	1
		K	0.93	0.52	1
		M	0.87	0.06	1
	Yahoo	P	0.89	0.27	1
		K	0.92	0.40	1
		M	0.88	0.24	1
CS	Google	P	0.94	0.52	1
		K	0.93	0.58	0.99
		M	0.86	0.42	0.99
	MSN	P	0.95	0.43	1
		K	0.95	0.65	1
		M	0.93	0.06	1
	Yahoo	P	0.81	0.35	0.99
		K	0.84	0.41	0.99
		M	0.83	0.37	0.99

Figure 3: Distance between WUI and API results on day n compared to day n+offset.
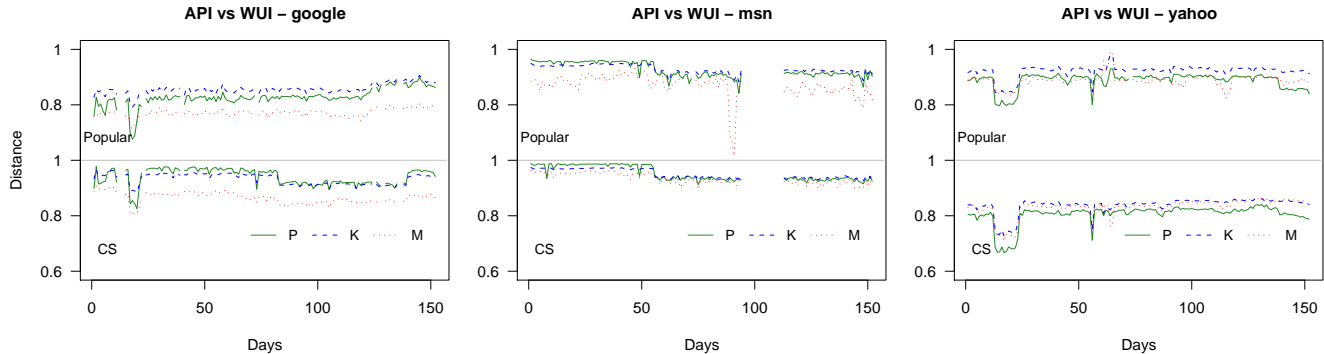
Figure 2: Distance between WUI and API top 100 search results.

Table 3: Identical WUI and API Search Results

	Identical in Rank		Identical in Membership	
	Top 100	Top 10	Top 100	Top 10
Google	0%	4.0%	3.1%	4.2%
MSN	0.2%	38.2%	6.4%	39.7%
Yahoo	0.2%	31.6%	0.2%	33.0%

ine only the top 10 results, Google's interfaces only give an identical result set 4% of the time although MSN and Yahoo improve considerably. There is only modest improvement by all of the search engines when we disregard ranking. In light of these findings, researchers should expect that the results obtained through any search engine's API will rarely be identical to what the general WUI user sees.

4.2.3 Decay of Search Results

A number of studies (e.g., [28, 36]) have examined the decay (or linkrot) of the Web over time. In a similar vein, we now examine how the search results for the same query decays over time. Search result decay is a previously unexamined phenomena. Although researchers have noted changes in search results over time, they have not reported change measures on a daily basis over a significant amount of time.

The decay of search results are influenced by a number of factors: 1) dynamicism of the Web (many web pages are here today and gone tomorrow), 2) changes to the content of web pages that make them more or less relevant to a given search term, 3) changes in page popularity (often measured by search engines in terms of inlinks), 4) indexing of new content by search engines, 5) de-duping and de-spamming efforts, and 6) changes to the search engine's ranking strategy. Although there are many factors influencing decay, we would expect each search engine's WUI and API to report similar decay rates. We would also expect the results for popular search terms to decay faster than those of common CS terms since we would expect new Web content in blogs, news websites, commercial sites, etc. to focus on popular subjects of the day.

To compute the decay of search results over time, we used the overlap distance measure P since it ignores rank. We have also computed the **search result half-life** for each search engine and each interface, that is the amount of time

it takes for half of the top k search results obtained on day n to completely drop from the result set.

There are two methods that may be used to compute the decay of search results over time. The first method is to compare the results on day 1 with each subsequent result obtained in our experiment. While this captures the specific changes for the results obtained on day 1, we are likely to see different changes over time if we had arbitrarily started with the results on day 10 instead, since there could have been a huge discrepancy between results on days 1 and 2 and a very small discrepancy between days 10 and 11. In order to remove this bias and calculate a more generalized set of changes over time, we use a second method whereby we averaged the offset results over the set of all results obtained in the experiment. For example, to calculate the decay for results that are 20 days apart, we averaged P between days (1,21), (2,22), ..., (n-20,n) where n is the total number of days in the experiment (150 in our case). More formally, the decay D for results d days apart is defined:

$$D(d) = \frac{\sum_{i=1}^{n-d} P(\tau_i, \tau_{i+d})}{n-d} \qquad (6)$$

When we examined the top 100 results for decay, each individual term produced results that decayed at different rates for each search engine. The interfaces reported nearly identical decay rates for each search engine except in a few cases like *jessica simpson* for Google and *programming language semantics* for Yahoo. When we examined only the top 10 results, the decay lines became somewhat less synchronized for the interfaces.

We also saw some results that "de-decayed" over time. For example, Google's WUI results for *mother's day* had decayed by almost half two months into the experiment but then rebounded to about 0.8 by the end of the experiment. On the other hand, Google's API reported a more linear decay line for *mother's day*.

To get a better composite picture of the decay for CS and popular term results, we averaged the decay rates for terms in each category. Figure 4 plots the decay of the top 100 search results over time for each of the three search engines. An **X** marks the half-life for each interface. For top 100 results using CS terms, we did not collected enough data to

Table 4: Half-Life of Top k Search Results

Type	Top	Google		MSN		Yahoo	
		WUI	API	WUI	API	WUI	API
Popular	100	94	71	69	73	47	48
	10	376*	529*	74	96	783*	433*
CS	100	215*	235*	327*	338*	44	48
	10	672*	1480*	228*	264*	67	71

* Values predicted by our model.

Table 5: 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%

observe a decay of 0.5 for Google and MSN. Nor did we observe a half-life for top 10 results except MSN using popular terms and Yahoo using CS terms. To predict these missing half-lives, we fitted the decay lines using the linear model with the highest R-squared values:

$$f(day) = a - b \cdot log(day) \qquad (7)$$

Variable a ranged from 0.912 to 1.167 (mean of 1.034), b from 0.161 to 0.393 (mean of 0.253) and R-squared from 0.951 to 0.994 (mean of 0.968). We refer the reader to [34] for specific values of a, b and R-squared. We then calculated the half-life by solving for day when $f(day) = 0.5$.

The observed and predicted half-lives for each search engine is presented in Table 4. For top 100 results, the predicted half-lives for both interfaces lie close to each other. For top 10 results, the predicted half-lives vary greatly; although we predict Google's top 10 half-life for CS terms at around 2 years for the WUI, the half-life is approximately 4 years for the API.

The CS results appear to be the most stable in Google and MSN. The most unstable results are from Yahoo whose popular and CS top 100 results have half-lives ranging from 44–48 days. In most cases the stability of the search results increased when examining only the top 10 results (MSN's CS results are the lone exception). This suggests that it typically takes much longer for a relevant search result to break into the first page of a search engine's results than it takes to break into the top 100.

4.3 Total Results per Search Term

The total results produced by a search term is a very rough estimate of the number of indexed pages containing the term since providing an accurate number is resource-intensive. Large numbers are usually impossible to verify since search engines will only return the first 1000 or so results whether using the API or WUI. Therefore we do not attempt to prove the accuracy of this number; we simply compare the numbers produced by the two interfaces.

Very rarely did any search engine's interfaces report the exact same total number of results. Google reported identical numbers about 2% of the time and Yahoo about 0.5% of the time. MSN consistently provided the exact same number (over 99% of the time) until July 21, 2006, when their interfaces began to report slightly different numbers. Since then, MSN's interfaces have agreed only 6% of the time.

Since the search engine interfaces frequently disagree, we define a more general sense of disagreement. *Loose disagreement* occurs when the API value is greater than or less than ± 10% of the WUI value. In Table 5 we summarize the percentage of loose disagreements for three of the query types we generated. In Figure 5 we have plotted the percentage of loose disagreements per day between the search engine

interfaces. We divide loose disagreements into two types, where the WUI reports higher or lower values than the API. The graphs in Figure 5 show that both Google and Yahoo typically report higher total result values for the WUI than the API. We also see the total percentage of disagreements between Google's interfaces growing from 40% to 70% on day 96 and then to 90% three weeks later. MSN's interfaces produce very little disagreement, and Yahoo's disagree consistently about 40% of the time.

4.4 Total Backlinks

In Figure 6 we have plotted the percentage of loose disagreements each day when examining the number of backlinks reported by each interface. Google's WUI and API typically agreed on every backlink request. For the most part, Google reported the same number of backlinks everyday for all 100 URLs, updating their values only occasionally. MSN and Yahoo were far less synchronized. Yahoo's disagreements held steady at about 35% until they began providing WUI responses from the Yahoo Site Explorer on day 72 (after day 72, Yahoo's standard WUI would always forward 'link:' and 'site:' requests to their new Site Explorer WUI).

Google reported backlink values that were unusually low when compared to the values reported by MSN and Yahoo. This is not a surprising finding since Google has acknowledged that they do not disclose all backlinks [15].

4.5 Total Pages Indexed per Website

Figure 7 plots the percentage of loose disagreements each day when examining the number of pages indexed as reported by each interface. Google appears to have had a major internal change on day 65 which caused more disagreement between the interfaces. Google's WUI regularly reports much larger values than the API. MSN appears to be the most consistent at returning identical values. Yahoo's increase in disagreements can again be blamed on the redirection to Yahoo Site Explorer beginning on day 72.

4.6 Indexed and Cached URLs

All three search engines appeared to give more consistent responses between their interfaces when asked about the indexed and cached status of URLs. Table 6 shows the percentage of times the WUI responded 'yes' to the queries 'Is URL x indexed?' and 'Is URL x cached?'. The table also reports the total number of URLs for which a disagreement was found.

Yahoo appears to have the greatest inconsistency between interfaces, averaging 7% disagreements when asked if a URL is indexed and cached. MSN and Google average fewer than 2% disagreements each day. Google only recorded disagree-
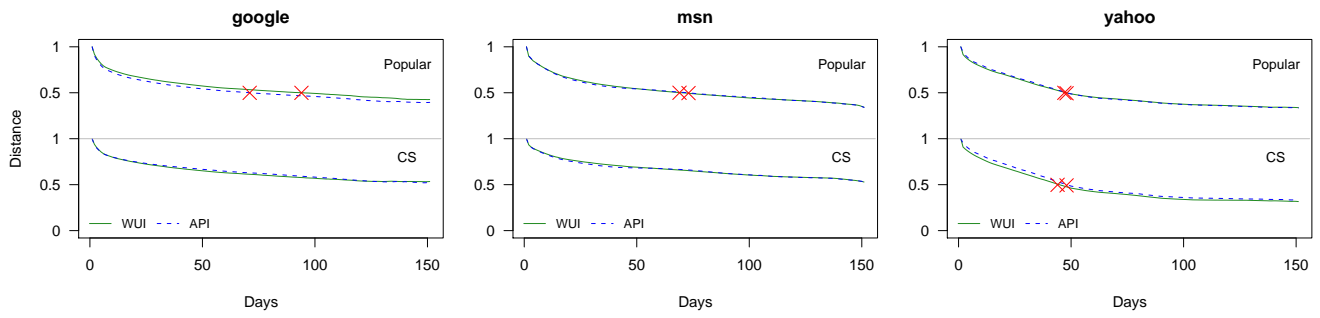
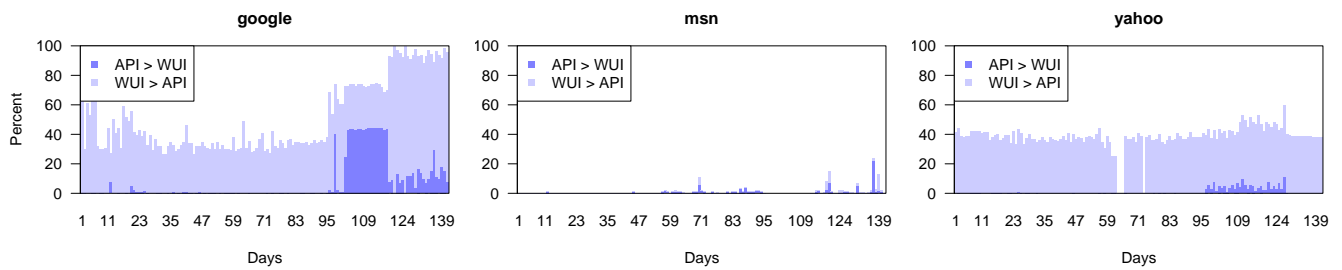Figure 4: Observed decay of top 100 search results over time and half-lives (X).

Figure 5: Daily percentage of loose disagreements between interfaces for total search results.
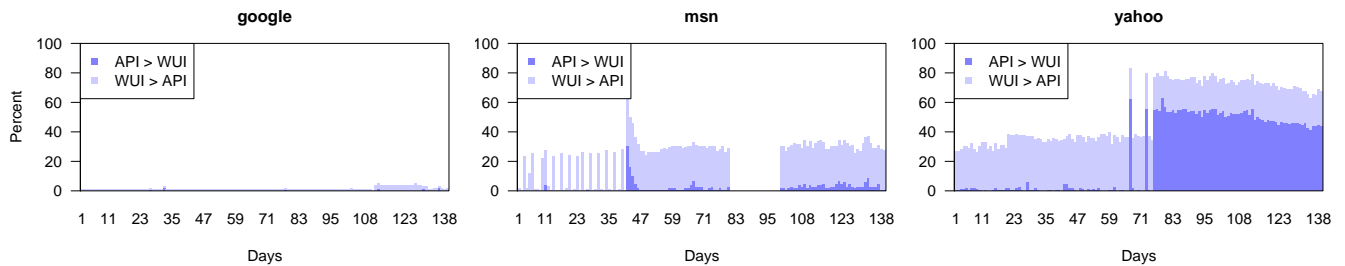

Figure 6: Daily percentage of loose disagreements between interfaces for total backlinks.
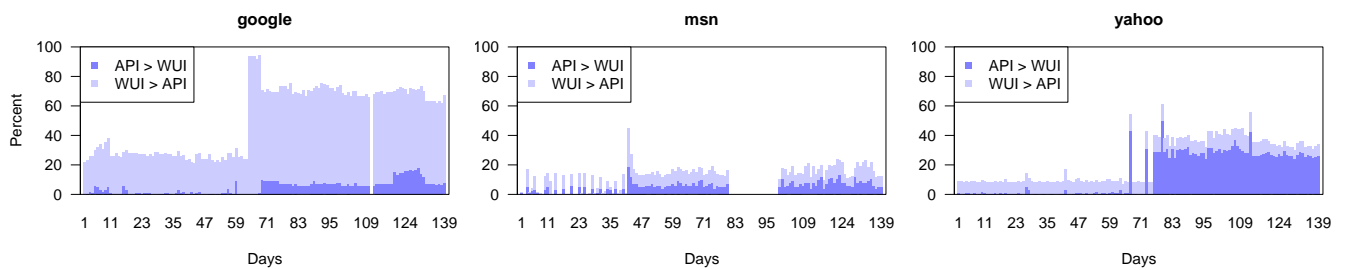

Figure 7: Daily percentage of loose disagreements between interfaces for total indexed pages.

Table 6: Indexed and Cached Status Disagreements

		Indexed		Cached	
		Disag	Total URLs	Disag	Total URLs
Google	WUI is yes	0.96%	4	1.08%	13
	API is yes	0.04%	6	0.69%	15
MSN	WUI is yes	0.24%	21	0.24%	21
	API is yes	0.91%	23	0.91%	23
Yahoo	WUI is yes	6.30%	28	6.52%	34
	API is yes	0.50%	29	0.49%	27

Table 7: Synchronized Interfaces

Type	Most synchronized	Least synchronized
Searching for popular terms	MSN	Google
Searching 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

ments for a handful of URLs and consistently reported `http://www.koolcelebrities.com/actress/angelina/` as being indexed by the WUI but not the API. MSN and Yahoo had trouble with a broader range of URLs.

5. DISCUSSION

Our analysis of search engine results over a five month span has shown that the there are significant differences between the reported values of the API and WUI interfaces. We can categorize those search engines which we found to have the least synchronized and most synchronized interfaces by examining the total loose disagreements and average distances between interfaces. In Table 7 we have classified the three search engines across all the query types we have examined. MSN appears to have the most synchronized interfaces overall. Although Yahoo provided the least synchronized results in terms of total backlinks, Yahoo has designed Site Explorer APIs that would likely be more synchronized than the Site Explorer WUI. We did not test these APIs because we were attempting to measure Yahoo's standard WUI to their standard API. Our findings show that Yahoo's standard API is working off an entirely different set of data than the Site Explorer WUI.

As noted in Section 4.2.2, a surprising result from our study was that the WUI and API interfaces typically give less similar results depending on the type of search term used. Popular terms revealed significantly lower distance measures between the interfaces for Google, and CS terms revealed significantly lower distance measures for Yahoo. Because the WUI and API results appear to be changing in a synchronized fashion, it appears that there are slight differences in ranking algorithms between the two which likely consider web spam differently.

When we ask the question, "Are the APIs serving results from an *older* index?", the answer is *no*. We have seen that all three search engines provide nearly synchronized changes in their results each day. Although the WUI and API results are consistently less than unity with regards to their distance measures, we found the highest distance measures always occurred on the same day (Figure 3).

When we ask the question, "Are the APIs serving results from a *smaller* index?", the answer is *probably* for Google and Yahoo. Google and Yahoo's WUIs consistently report total results that are higher than the APIs, but MSN's interfaces regularly agree (of course we cannot directly verify these estimates, and Google adds 'supplemental results' to their WUI results). The backlink counts from MSN's and Yahoo's WUI were also consistently larger than the API counts. And Google's WUI consistently produced larger website page counts than did their API. To give a more definitive answer to this question, we suggest a future experiment that randomly samples from each corpus and compares the overlap [8] or one that uses other methods [12].

6. CONCLUSIONS

Our five month experiment has uncovered a variety of disagreements between the interfaces of Google, MSN and Yahoo. Our findings suggest that the API indexes are not older, but they are *probably* smaller. Although the indexes used by the WUI and API appear to be updated at the same rate for all three search engines, the top 100 WUI and API results are rarely identical for CS and popular term queries. When examining just the top 10 results, Google's API produces results that are 20% different than the WUI results; Yahoo's are 14% different, and MSN's are 8% different. In general, we found MSN to produce the most consistent results between their two interfaces.

We have also examined how search results decay over time. We have built predictive models based on the observed decay rates of the popular and CS search term results used in our experiment. In general, we have found that the decay rates for popular results differ significantly for Google and MSN, and the top 10 results decay at a much slower rate than do the top 100 results. It can take over a year for half of the results to a popular query to be replaced with other results in Google and Yahoo; for MSN it can take only 2-3 months.

We hope that our findings will allow other researchers to better understand the differences between results obtained through the search engine interfaces. Researchers may need to use caution when generalizing their results obtained from the API to those results that the common user sees using the WUI. 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.

On December 20, 2006, Google announced they were beginning to deprecate their SOAP API in favor of an AJAX API [13]. It remains to be seen if the AJAX API will exhibit a higher of level of synchronization with the WUI.

7. REFERENCES

[1] Ask terms of service, 2006. `http://sp.ask.com/en/docs/about/terms_of_service.shtml`.
[2] J. Bar-Ilan. Search engine results over time - A case study on search engine stability. *Cybermetrics*, 2/3(1), 1998/99.
[3] J. Bar-Ilan. Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology*, 53(4):308–319, 2002.
[4] J. Bar-Ilan. Comparing rankings of search results on the Web. *Information Processing & Management*, 41(6):1511–1519, Dec. 2005.
[5] J. Bar-Ilan. Expectations versus reality – search engine features needed for web research at mid 2005. *Cybermetrics*, 9(1), 2005.

[6] J. Bar-Ilan, M. Levene, and M. Mat-Hassan. Dynamics of search engine rankings - A case study. In *Proceedings of the 3rd International Workshop on Web Dynamics*, May 2004.

[7] 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.

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

[9] D. Bergmark. Collection synthesis. In *Proceedings of JCDL '02*, pages 253–262, 2002.

[10] D. Bergmark, C. Lagoze, and A. Sbityakov. Focused crawls, tunneling, and digital libraries. In *Proceedings of ECDL '02*, pages 91–106, 2002.

[11] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of WWW7*, pages 379–388, 1998.

[12] A. Broder, M. Fontura, V. Josifovski, R. Kumar, R. Motwani, S. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu. Estimating corpus size via queries. In *Proceedings of CIKM '06*, pages 594–603, 2006.

[13] D. Clinton. Beyond the SOAP search API, Dec. 2006. `http://google-code-updates.blogspot.com/2006/12/beyond-soap-search-api.html`.

[14] K. Curran and A. Doherty. Automated broadcast media monitoring using the Google API. In *Proceedings of CCNC 2006*, volume 2, pages 1098–1102, 2006.

[15] M. Cutts. GoogleGuy's posts, June 2005. `http://www.webmasterworld.com/forum30/29720.htm`.

[16] M. Cutts. Google datacenters. Video, July 31 2006. `http://video.google.com/videoplay?docid=8726665066825965913`.

[17] Did-it, Enquiro, and Eyetools uncover search's Golden Triangle, 2005. `http://www.enquiro.com/eye-tracking-pr.asp`.

[18] W. Ding and G. Marchionini. A comparative study of web search service performance. In *Proceedings of the ASIS Annual Meeting*, volume 33, pages 136–142, 1996.

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

[20] P. Festa. Google worm targets AOL, Yahoo. Dec. 28 2004. `http://news.com.com/Google+worm+targets+AOL%2C+Yahoo/2100-7349_3-5504769.html`.

[21] S. Gauch, G. Wang, and M. Gomez. Profusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computer Science*, 2(9):637–649, 1996.

[22] B. Gillette. Google blacklisting researchers? Dec. 14 2004. `http://www.emailbattles.com/2005/12/14/virus_aacdehdcic_ei/`.

[23] Google privacy center: Terms of service, 2006. `http://www.google.com/terms_of_service.html`.

[24] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Proceedings of WWW '05*, pages 902–903, May 2005.

[25] T. G. Habing, T. W. Cole, and W. H. Mischo. Developing a technical registry of OAI data providers. In *Proceedings of ECDL '04*, pages 400–410, 2004.

[26] N. Jain, M. Dahlin, and R. Tewari. Using Bloom filters to refine web search results. In *Proceedings of the 8th International Workshop on the Web and Databases*, 2005.

[27] M. Klein, M. L. Nelson, and J. Z. Pao. Augmenting OAI-PMH repository holdings using search engine APIs. In *Proceedings of JCDL '07*, 2007.

[28] W. Koehler. A longitudinal study of web pages continued: A consideration of document persistence. *Information Research*, 9(2), 2004.

[29] M. Koo and H. Skinner. Improving web searches: Case study of quit-smoking web sites for teenagers. *Journal of Medical Internet Research*, 5(4), Nov. 2003.

[30] R. Kraft and R. Stata. Finding buying guides with a web carnivore. In *First Latin American Web Congress (LA-WEB'03)*, pages 84–92, 2003.

[31] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Intelligence*, 11(1):32–39, 2000.

[32] The Lycos 50, 2006. `http://50.lycos.com/`.

[33] P. Mayr and F. Tosques. Google Web APIs - an instrument for webometric analyses? In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI '05)*, 2005.

[34] F. McCown. Comparison of search engine interfaces, 2006. `http://www.cs.odu.edu/~fmccown/research/se_apis/`.

[35] F. McCown, J. Bollen, and M. L. Nelson. Evaluation of the NSDL and Google for obtaining pedagogical resources. In *Proceedings of ECDL '05*, pages 344–355, 2005.

[36] F. McCown, S. Chan, M. L. Nelson, and J. Bollen. The availability and persistence of web references in D-Lib Magazine. In *Proceedings of the 5th International Web Archiving Workshop (IWAW '05)*, Sept. 2005.

[37] F. McCown, X. Liu, M. L. Nelson, and M. Zubair. Search engine coverage of the OAI-PMH corpus. *IEEE Internet Computing*, 10(2):66–73, Mar/Apr 2006.

[38] M. Moffatt. Yahoo error: Unable to process request at this time – error 999. Feb. 14 2005. `http://murraymoffatt.com/software-problem-0011.html`.

[39] MSN terms of service, 2006. `http://tou.live.com/en-us/default.aspx`.

[40] MSN Web Search API. `http://msdn.microsoft.com/msn/msnsearch/`.

[41] G. Pant. Deriving link-context from HTML tag tree. In *Proceedings of DMKD '03*, pages 49–55, 2003.

[42] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles. Panorama: extending digital libraries with topical crawlers. In *Proceedings of JCDL '04*, pages 142–150, 2004.

[43] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the data: Parallel analysis with Sawzall. *Dynamic Grids and Worldwide Computing*, 13(4):277–298, Nov. 2005.

[44] C. Snelson. Sampling the Web: The development of a custom search tool for research. *Library and Information Science Research Electronic Journal*, 16(1), Dec. 2005.

[45] A. Spink, B. J. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Information Processing & Management*, 42(5):1379–1391, Sept. 2006.

[46] K. C. Sua, S. E. Waldren, and T. B. Patrick. Differences in the effects of filters on health information retrieval from the internet in three languages from three countries: A comparative study. In *Proceedings of MEDINFO 2004*, 2004.

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

[48] L. Vaughan. New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, 40(4):677–691, May 2004.

[49] What's a "supplemental result?". *Google Webmaster Help Center*, 2006. `http://www.google.com/support/webmasters/bin/answer.py?answer=34473`.

[50] Wikipedia: List of basic computer science topics, 2006. `http://en.wikipedia.org/wiki/List_of_basic_computer_science_topics`.

[51] Yahoo! Web Search APIs. `http://developer.yahoo.net/search/web/`.

[52] Z. Zhuang, R. Wagle, and L. C. Giles. What's there and what's not?: Focused crawling for missing documents in digital libraries. In *Proceedings of JCDL '05*, pages 301–310, 2005.