Search Engine Retrieval of Changing Information

Yang Sok Kim
University of Tasmania
School of Computing, University of Tasmania, Private Bag 100 Hobart
TAS 7001 Australia
+61 3 6226 2907
yangsokk@utas.edu.au

Byeong Ho Kang
University of Tasmania
School of Computing, University of Tasmania, Private Bag 100 Hobart
TAS 7001 Australia
+61 3 6226 2919
bhkang@utas.edu.au

Paul Compton
The University of New South Wales
School of Computer Science and Engineering
The University of New South Wales
Sydney 2052 Australia
compton@cse.unsw.edu. au

Hiroshi Motoda
Osaka University
The Institute of Scientific and Industrial Research, Osaka University, 8-1
Mihogaoka, Japan
motoda@sanken.osaka-u.ac.jp

ABSTRACT

In this paper we analyse the Web coverage of three search engines, Google, Yahoo and MSN. We conducted a 15 month study collecting 15,770 Web content or information pages linked from 260 Australian federal and local government Web pages. The key feature of this domain is that new information pages are constantly added but the 260 web pages tend to provide links only to the more recently added information pages. Search engines list only some of the information pages and their coverage varies from month to month. Meta-search engines do little to improve coverage of information pages, because the problem is not the size of web coverage, but the frequency with which information is updated. We conclude that organizations such as governments which post important information on the Web cannot rely on all relevant pages being found with conventional search engines, and need to consider other strategies to ensure important information can be found.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/ Hypermedia; K.4.m [**Computers and Society**]: Miscellaneous; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Measurement, Performance, Design

Keywords

Web coverage, Web characterization, Web monitoring, rate of change, overlap of Web search result, search engines

1. INTRODUCTION

Originally the Web was designed to support passive and distributed information delivery using a receiver-pull model. In this model, information publishers post material to the Web without any notification to potential readers; readers are expected to visit the website if they want the information. Although this approach gives advantages such as protection from unwanted traffic (e.g e-mail spam)[1] and ad hoc management[1][2], it does not guarantee that the published information is found and used by those for who it is intended. For this reason, methods for finding information have been heavily researched since the beginning of the Web.

Web crawling and the Web monitoring are different methods to support the Web information finding. A Web crawler program starts with given a set of URLs. It visits all the URLs and extracts any URLs from the visited Web pages which are collected. It may then repeat the process, visiting these URLs, until the crawler decides to stop, for any one of a number of reasons. Nowadays crawlers are widely used by the major search engines (e.g., Google, AltaVista, and Excite) to collect a significant number of textual Web pages [3]. A crawler program will revisit web pages, to check if the links on the page have changed, but the emphasis is on moving out, following links to find as many pages as possible, rather than revisiting. A Web monitor program also starts with a given set of URLs. However, it does not follow the URLs to find new URLs; rather, it continually revisits the given URLs and compares the old Webpage URLs with the new Webpage URLs to extract newly created URLs [4-7].

It is useful to distinguish a **Web information source pages (P_s)** from a **Web information pages (P_i)**. Though the latter may contain URLs to the other Web page, their main purpose is to display a specific content. On the contrary, the former are Web pages mainly used to provide URLs to Web information pages. For example, let us assume that you visit an online newspaper Web site such as CNN or the BBC. When you visit the Web site, it usually displays a list of news articles that are linked to content pages. According to our definition the list page is a Web information source page and the content page is a Web information page. From this definition, the key difference between the Web crawlers and the Web monitors can be characterised as follows: Whereas Web crawlers try to find both new P_s and new P_i starting from the seed pages, Web monitors find only new P_i from the given P_s.

WWW 2007, May 8--12, 2007, Banff, Canada.

[1] Ad hoc management is typical of transient tasks: you connect to a network device, retrieve some data to check something, and disconnect shortly after.

2. PROBLEMS

Web search engines have significantly different performance because they use different crawling and indexing strategies. Their performance can be measured by their **coverage** (how many Web pages the system covers from the entire Web), their **relevance** (how precisely the system provides the user with information that is relevant to their interests), and their **timeliness** (how up to date the system is with new information).

(1) Coverage: Although search engine providers have continually competed to expand their coverage, previous research results show that the current coverage of each search engine is significantly different [8-10] and the entire coverage of all search engines is only a fraction of the entire Web [11]. We studied the coverage problem by comparing crawling results with monitoring results assuming that a web monitor would go closer to collecting all the new information pages from given Web information source pages ($\boldsymbol{P_s}$), than a crawler. We compared coverage of the information pages found by our Web monitor program with the coverage of these pages by Google, Yahoo, and MSN.

(2) Relevance: It is critical for all search engines to provide relevant information simply because of the vast amount of information available. Simon[12] commented: "What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." Indexing and ranking algorithms contribute to the performance of Web search engines. Our study does not directly address this critical issue.

(3) Timeliness: Web search engines aim to provide fresh information. Search engines suffer from having to make a trade-off between timely information and coverage.. The greater the coverage of the search engine, the longer it takes to revisit all pages. It can be important to know the delay between information creation time ($\boldsymbol{T_c}$) by the publisher and information recognition time ($\boldsymbol{T_r}$) by the crawler. This delay ($\boldsymbol{D}$) is defined as $\boldsymbol{T_c}$ - $\boldsymbol{T_r}$. Sometimes this delay causes some Web information pages to be missed because their URLs are removed from their related information source pages to provide space for links to new information pages before the crawlers revisit.

In this paper, we compare the coverage, overlap, and dominance of three well-known commercial search engines on information pages found by our Web monitor program. By overlap, we mean coverage of the same pages by different search engines, and by dominance, we simply mean whether one search engine's coverage is much better than the others. Section 3 outlines our research methodology, and describes Web monitoring, Web information source page selection, monitoring scheduling strategies and the data set which is collected from Australian Federal/Local Government homepages and media release pages. Section 4 presents the coverage analysis results. Section 5 presents overlap and dominance results. Section 6 discusses these results and Section 7 presents the conclusions the study.

3. EVALUATION METHODOLOGY

In order to evaluate coverage, we need to sample Web pages and match the contents of these with the results returned from search engines. The following two sections outline how to sample Web pages and how to collect information from these Web pages.

3.1 Information Source Pages Sampling Strategy

In this type of study, the selection of sampling sites is crucial. Many studies use randomly selected samples [3, 10, 13-16]. However, scaling to the entire Web is still not clear because no one knows the boundary of the Web. We consider two characteristics of Web pages, reach-ability by crawlers and frequency of content updating. We selected 260 Australian Government Web pages (Table 1) including both homepages for various departments and media release pages. The Local Government web pages include web pages from both the Tasmanian State Government and municipal government services in Tasmania, thus accounting for the higher number of homepages and smaller number of media release pages compared to the Federal Government. Obviously this sample set will not test the overall performance of Web search engines but we believe that they are not extreme cases with respect to reach-ability by crawlers and frequency of content updating.

Table 1. Web Information Source Pages by Domain

		Number of pages monitored	Ratio
Federal Government	homepages	14	5%
	Media release pages	118	45%
Local Government	homepages	111	43%
	Media release pages	17	7%
	Total	260	100%

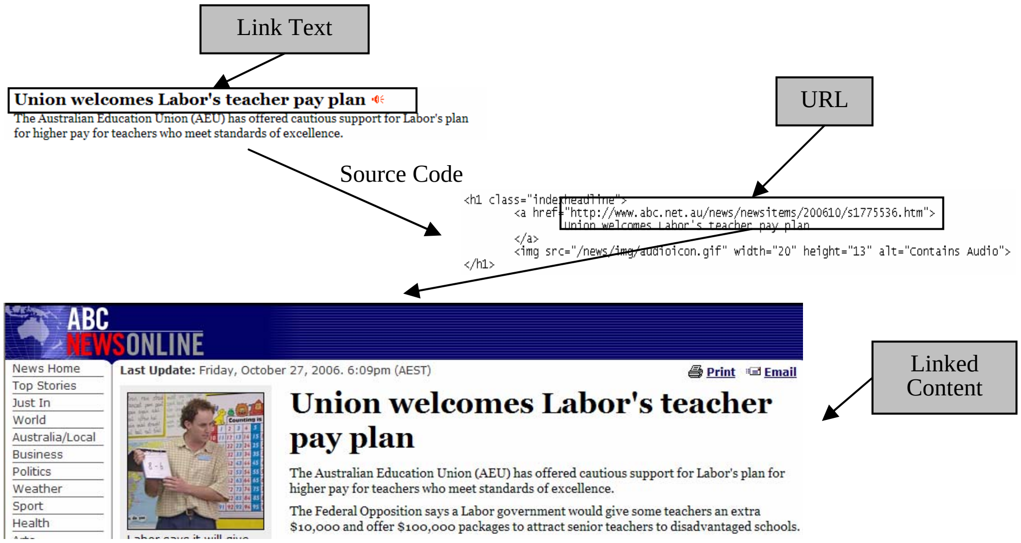
3.2 Collection of Datasets from Sample Web

3.2.1 Data Collection System

The Web monitor program, WebMon has been described previously[17]. As shown in Figure 1 we use the terms: "Link text" is located between <a> and </a>tags and displayed in a Web browser. "URL" indicates the location of the content document. "Linked content" is the main content to be read by users. When the monitor program starts to run, it downloads the Web information source page and extracts all URLs (URL$_{old}$). At specified intervals, it revisits the Web page and repeats the process to get new URLs (URL$_{new}$). The monitor identifies new information pages ($\boldsymbol{P_i}$) by comparing URL$_{old}$ and URL$_{new}$ and eliminating filtered URLs (URL$_{filter}$)[2]. For each information page the URL, link text, and linked content are stored for further processing and URL$_{new}$ becomes URL$_{old}$. This process is repeated

[2] WebMon includes a list of advertisement URLs, which is used to filter out advertisements. The list is added to when new sources of advertising are discovered and is reasonably complete for the type of pages studied here

indefinitely. This monitoring process and change detection criteria are illustrated in Figure 2.



Union welcomes Labor's teacher pay plan
The Australian Education Union (AEU) has offered cautious support for Labor's plan for higher pay for teachers who meet standards of excellence.

```
<h1 class="indexheadline">
        <a href="http://www.abc.net.au/news/newsitems/200610/s1775536.htm">
        Union welcomes Labor's teacher pay plan
        </a>
        <img src="/news/img/audioicon.gif" width="20" height="13" alt="Contains Audio">
</h1>
```

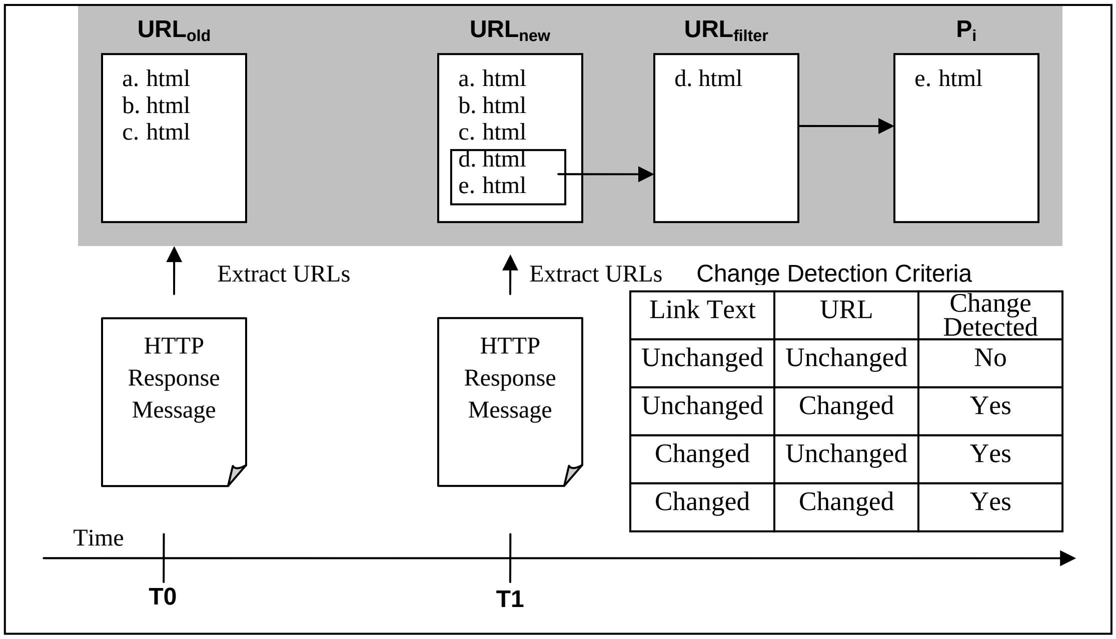
Last Update: Friday, October 27, 2006. 6:09pm (AEST)

Union welcomes Labor's teacher pay plan

The Australian Education Union (AEU) has offered cautious support for Labor's plan for higher pay for teachers who meet standards of excellence.

The Federal Opposition says a Labor government would give some teachers an extra $10,000 and offer $100,000 packages to attract senior teachers to disadvantaged schools.

Figure 1. Link Text, URL and Linked Content



Change Detection Criteria

Link Text	URL	Change Detected
Unchanged	Unchanged	No
Unchanged	Changed	Yes
Changed	Unchanged	Yes
Changed	Changed	Yes

Figure 2. Monitoring Process and Criteria for Detecting Changes

3.2.2 Monitoring Scheduling Strategy

The Revisiting time ($T_{revisit}$) set for monitoring depends on the publication frequency of the source pages[3]. If the publication frequency is high, $T_{revisit}$ should decrease otherwise increase. There are three different strategies for monitoring scheduling.

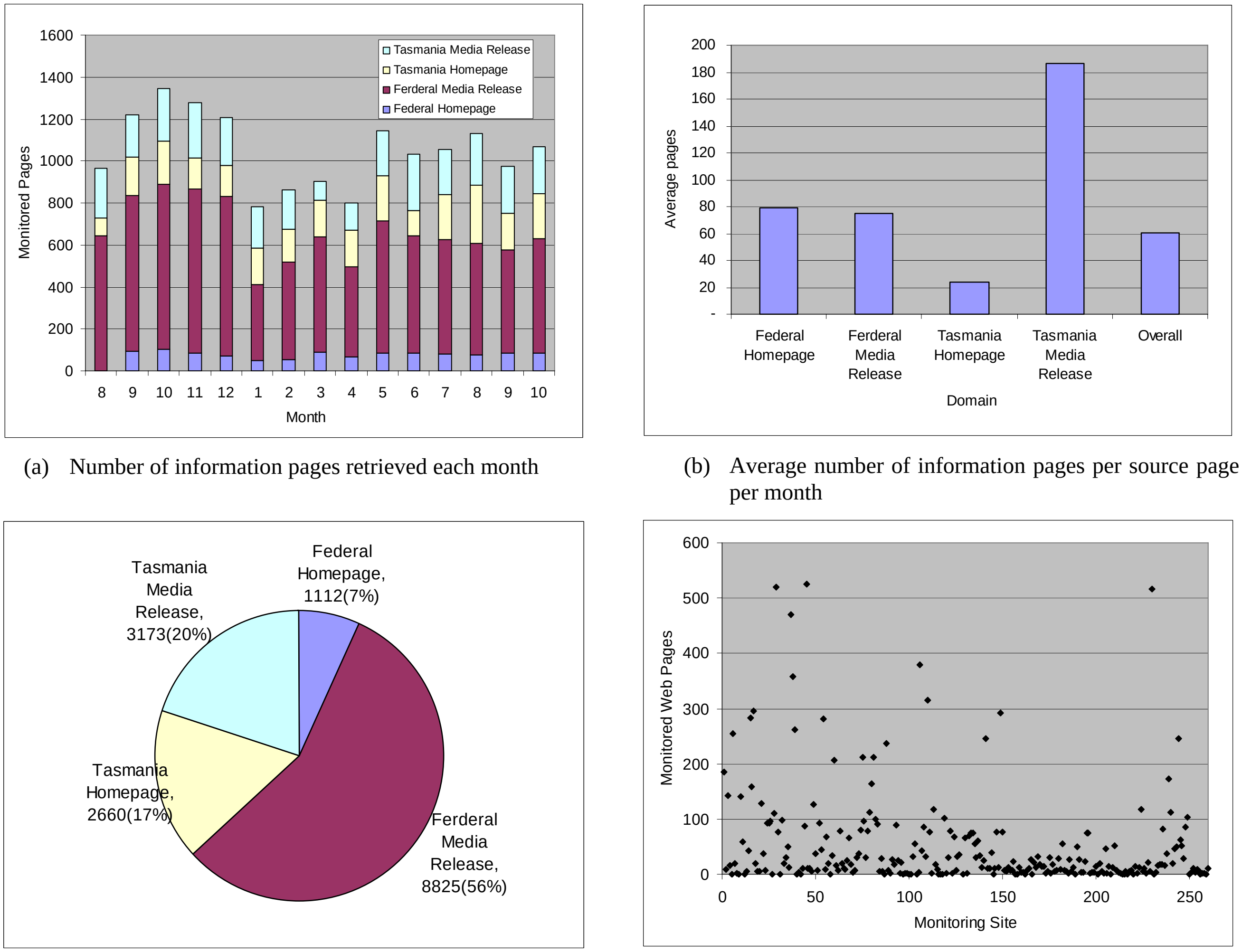
(1) Single fixed scheduling strategy: $T_{revisit}$ is set as a fixed interval such as every 30 minutes, every day 9:00 am, or 9:00 am every Monday. This does not consider the publication frequency and user’s needs. This strategy would be employed when the monitor program has no prior information.

(2) Multiple fixed scheduling strategy: $T_{revisit}$ is set as a fixed time individually for each source page.

(3) Dynamic and adaptive scheduling strategy: In this strategy $T_{revisit}$ is dynamically changed to reflect publication frequency. There is a trade-off between the cost of monitoring and the cost of missing information.

Although (2) and (3) are more efficient and cost-effective, the fixed scheduling strategy was used here as we had no prior information about the publication patterns of the selected domains and the focus of the research is on analysing the performance of Web crawlers using Web monitoring rather than strategies for web monitoring. We set the revisit time ($T_{revisit}$) as 2 hours for all the Web information source pages.

[3] In the Web monitoring system, the way in which the user is notified is also factor for scheduling, but is not relevant to the study here.



(a) Number of information pages retrieved each month

(b) Average number of information pages per source page per month

(c) Total retrieval by area

(d) Average number of information pages per individual source page per month

Figure 3. Analysis on collected data set from sample pages

3.2.3 Analysis of Collected Dataset

We collected new Web information pages from August 2005 to October 2006. In total 15,770 new Web information pages (Data set)[4] were collect from the 260 Web information source pages (Sample Pages). These are public web pages which should be readily accessible to any web crawler.

Figure 3 (a) illustrates monthly monitoring trends during the monitoring period. On average 1,051±175.5 Web information pages per month were collected by WebMon. The maximum number was 1,344 during October, 2005 and the lowest number was 784 during January, 2006. Variations in the frequency of federal media releases are main factor in the overall variation. The average number of information pages published per month per Web Information source page is shown in Figure 3(b), with Tasmanian Media Release source pages being the most prolific, but note there are far fewer Tasmanian media release source pages than for Federal Government media releases, so that these results do not mean more Tasmanian media releases than Federal. One particular Tasmanian media release page for State Government media releases, is particularly anomalous with 2,468 information pages. Figure 3 (c) shows the total number of information pages retrieved by area and as might be expected federal media releases were the highest. Figure 3 (d) shows the number of web pages per month per information source page. There is wide variation and to avoid masking this, the Tasmanian government media release Web page, with 2,468 pages is not shown. Most source pages published fewer than 100 information pages during the monitoring period.

3.3 Query Method for Measuring Coverage

To check coverage by search engines, we do not simply retrieve the URL as the content may have changed. Rather we submit a query based on the contents of the information page and then check if the page is included amongst those retrieved. To do this we need to consider (1) what can be used as a query string, (2) how we decide the search engine returns contains the right result, and (3) how many returns we should check for this decision.

[4] We did not include Web information pages that have no 'link text' or that do not have not sufficient text for querying after removing stop words.

3.3.1 Formation of Query String

A query string can be extracted from the 'link text' or 'linked content'. In these experiments we use link text as the query string. We do not perform stop word elimination or stemming, because each search engine has different query processing methods. Firstly, the query string is submitted for exact matching, enclosed by quotation marks (e.g. "Support for young people in Wadeye"). If the search engine does not return any positive result, the query string is submitted without quotation marks.

3.3.2 Positive Result

For a '***positive result***' the URL of the information page must exactly matches one of the URLs returnd. Otherwise, it is called '***negative result***'. We define a positive result rate (PRR) as follows:

$$PRR(i) = \frac{R(i)}{S}$$

where *R(i)* is total positive return within *i* rank(s) and *S* is total number of sample. *PPR(i)* is an indicator of how well the search engines locate the relevant page.

3.3.3 Threshold for the Search Engine Results

Usually the default size of a search return is fixed at 10 or 20 pages. However, the relevant page may not occur within the first 10 or 20 pages, particularly if the query is very general. To evaluate retrieval we evaluated 375 random samples (confidence level 95%, confidence interval 5%) of 'link text' and checked the top 500 results. If there is a positive result, the location of the result is stored. The results show that 95.5 % of positive results are in the top 100 results (see Figure 4). For the rest of the study we considered only the first 100 returns.

3.3.4 Sampling Strategy for the Main Experiment

Random sampling from the entire data set was necessary in this evaluation because search engines constrain or monitor the number of automated searches by same user / IP. When the population size is infinite, the appropriate sample size of each month was calculated using:

$$S = \frac{Z^2 * (p) * (1-p)}{c^2}$$

, where

Z = Z value (e.g. 1.96 for 95% confidence level)

p = percentage picking a choice

c = confidence interval

When the population size is finite, the following formula is used to get the appropriate sample size.

$$\text{New S} = \frac{S}{1 + \frac{S-1}{P}}$$

, where S is the sample size of the infinite population and P is population size.

We sampled the data set as follows with 95% confidence levels and a 5% confidence interval. 4203 samples were selected, 23% of all monitoring results.

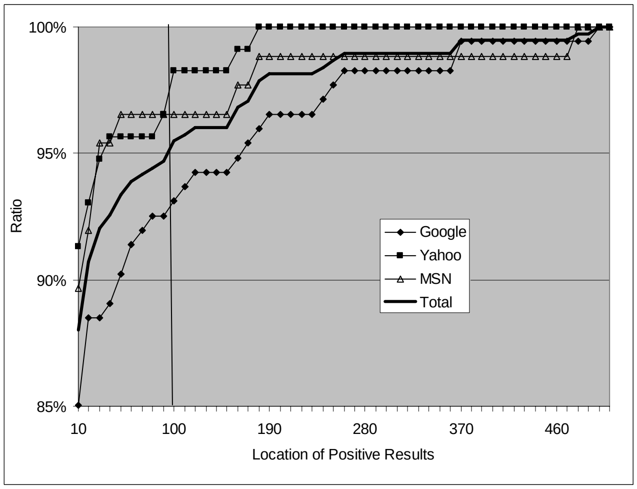


Figure 4. Number of Search Engines Results considered

Table 2 Sample Size (Confidence Level 95%, Confidence Interval 5%)

month	**8**	**9**	**10**	**11**	**12**	**1**	**2**	**3**
Monitored pages	964	1222	1344	1280	1205	784	861	905
Sample	275	292	299	296	291	258	266	270
month	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**total**
Monitored pages	800	1142	1032	1056	1130	975	1070	15770
Sample	260	288	280	282	287	276	283	4203

Table 3 Coverage Results

		Monitored Web Information Pages	Google		MSN		Yahoo	
			Positive Results	PRR	Positive Results	PRR	Positive Results	PRR
Federal	Home	289	153	52.9%	106	36.7%	87	30.1%
	Media	2,328	1,316	56.5%	700	30.1%	930	39.9%
Local	Home	724	258	35.6%	115	15.9%	135	18.6%
	Media	862	544	63.1%	32	3.7%	102	11.8%
Total		**4,203**	**2,271**	**54.0%**	**953**	**22.7%**	**1,245**	**29.8%**

4. COVERAGE ANALYSIS

4.1 Overall Coverage

Table 3 summarizes the overall coverage results for the three search engines. The coverage performance is the proportion of pages or positive result ratio (PRR) defined previously. Google gives the highest overall return and MSN the lowest. Overall Google returns 54.0% of the information pages and MSN 22.7%. That is they miss from 46.0% to 77.3% of the Web information pages that have been posted. The search engines also perform differently across different areas. For Google, local government media release pages give the best results, while local home pages give the worst return. In contrast for both MSN and Yahoo, local government media release pages give the worst results.

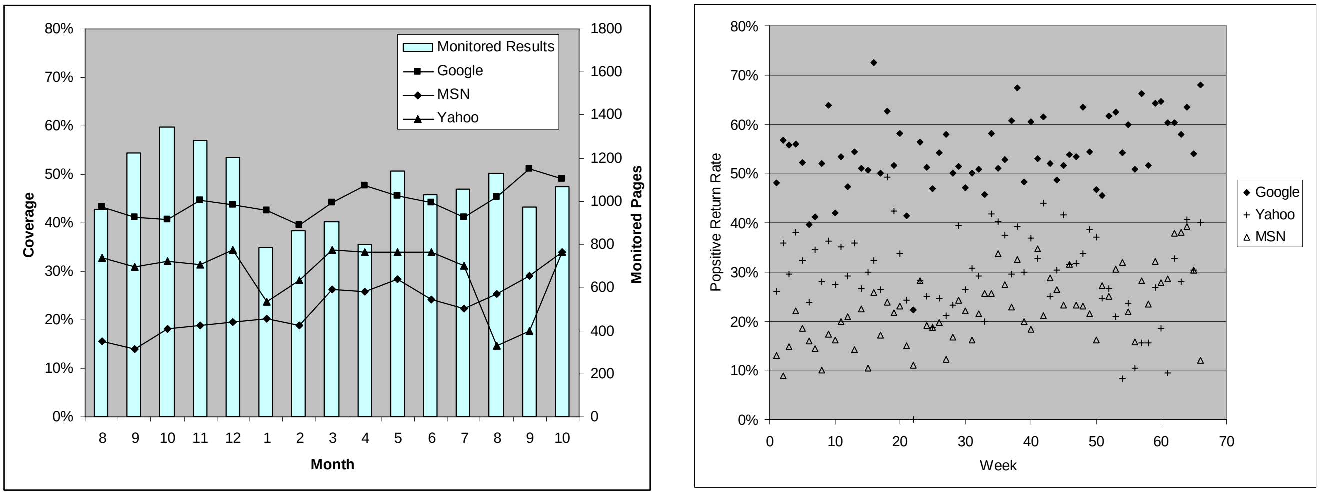
4.2 Coverage Trends

Figure 5 (a) illustrates coverage trends during the monitoring. We find following trend characteristics:

(1) As their with the summary results, the month by month results show that Google is consistently the best with Yahoo second, except for an anomalous period at the end, and MSN third.

(2) The week by week results in Figure 5 (b) show the variation in coverage performance more precisely. The variations in coverage are affected by factors such as the crawler's source page coverage, revisit schedule, the number of URLs that a source page can contain, and the number of publications.

(3) Google and MSN search engines broadly give higher returns in more recent months. This might have been because of improved crawling during the period, but is more likely that they might use crawled date or indexed date as one of results ranking factors. Yahoo does not improve over time, but the sudden change at the end suggests a possible changes to they way they crawl the Web.



(a) Coverage Trends

(b) Distribution of Coverage

Figure 5. Monitoring Results

Figure 6 illustrates coverage in the four different domains.

(1) Google outperformed other search engines in each month in all the domains. Especially, Google shows far better performance in the local government media release domain. As noted Yahoo coverage changes in the later months, with the most significant decline being in federal government homepage coverage and an extraordinary increase in local government media release coverage in the last two months.

(2) There appears to be greater fluctuation in the coverage of the homepage domains than for the media release domain. We conjecture that the crawlers visit the Web information source pages for the media release domains more frequently than for the homepage domain because these pages publish more Web information pages than the source pages for the homepage domain.

5. OVERLAP, DOMINANCE, AND UNIQUENESS ANALYSIS

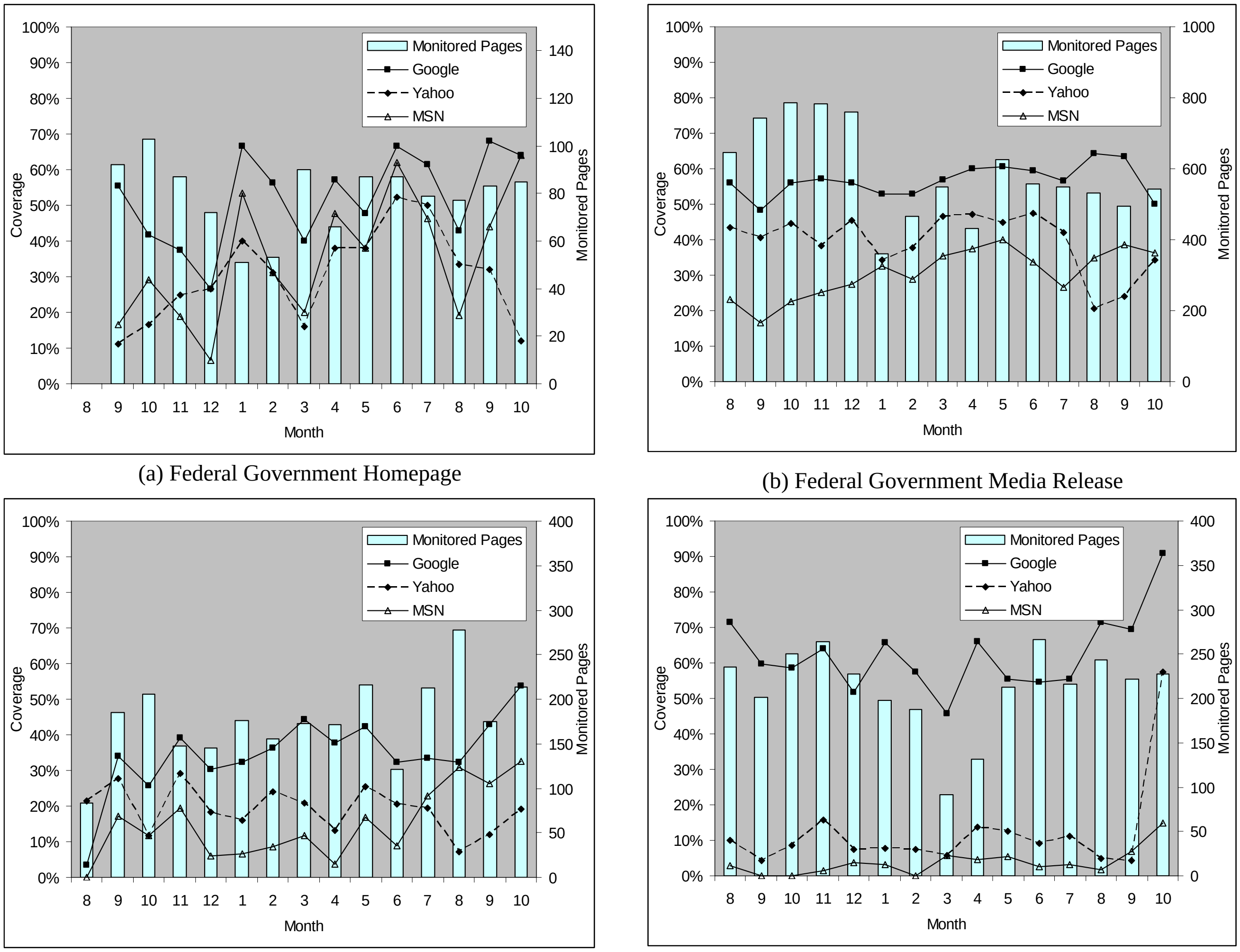
Overlap analysis was conducted, to see if the results suggested that a meta-search engine would produce better results by combining results for the three search engines. We found the following:

(1) The overlap of all three search engines' positive results is 9.7% (433/4,478 results) of the total returns (Google 2,271, MSN 953, Yahoo 1,254, see Table 3). Overlap ratios between pairs of search engines are as follows:
- Overlap ratio between Google and Yahoo: 974/(2,271+1,254) = 27%
- Overlap ratio between Google and MSN: 782/(2,271+953) = 24.3%
- Overlap ratio between MSN and Yahoo: 490/(953+1,254) = 22.2%

(2) Total unique positive returns (TUPR) are 2,665, 63.4% of the monitored Web information pages. It is calculated as follows:

TUPR=G(2,271)+Y(1,254)+M(953)–GM(782)–GY(974)–MY(490)+GMY(433), where G, Y, M, GM, GY, MY, and GMY represent positive results from Google and their overlapped positive returns (see Figure (a)).

(3) Google dominates the other search engines as shown in Figure 7(a) and Table 4 (a). 78% (974/1,254) of Yahoo's positive results are overlapped by Google and 82% (782/953) of MSN's positive results. In other words, whereas 42% (948 = G-GM-GY+GMY) of the Google's positive results are unique, only 12% (114=M-GM-MY+GMY) of MSN's and 18% (223=Y-GY-MY+GMY) of Yahoo's positive results are unique. This result does not suggest a significant improvement by using a meta search engine. Figure 7 (d) illustrates that Google's coverage of the unique positive results was about 85% during the monitoring period.



(a) Federal Government Homepage

(b) Federal Government Media Release

(c) Local Government Homepage

(d) Local Government Media Release

Figure 6. Monitoring Results by Domain

(4) Google's dominance is not constant over the monitoring period. We separate positive results as illustrated in Figure 7 (a), into the following three sections:

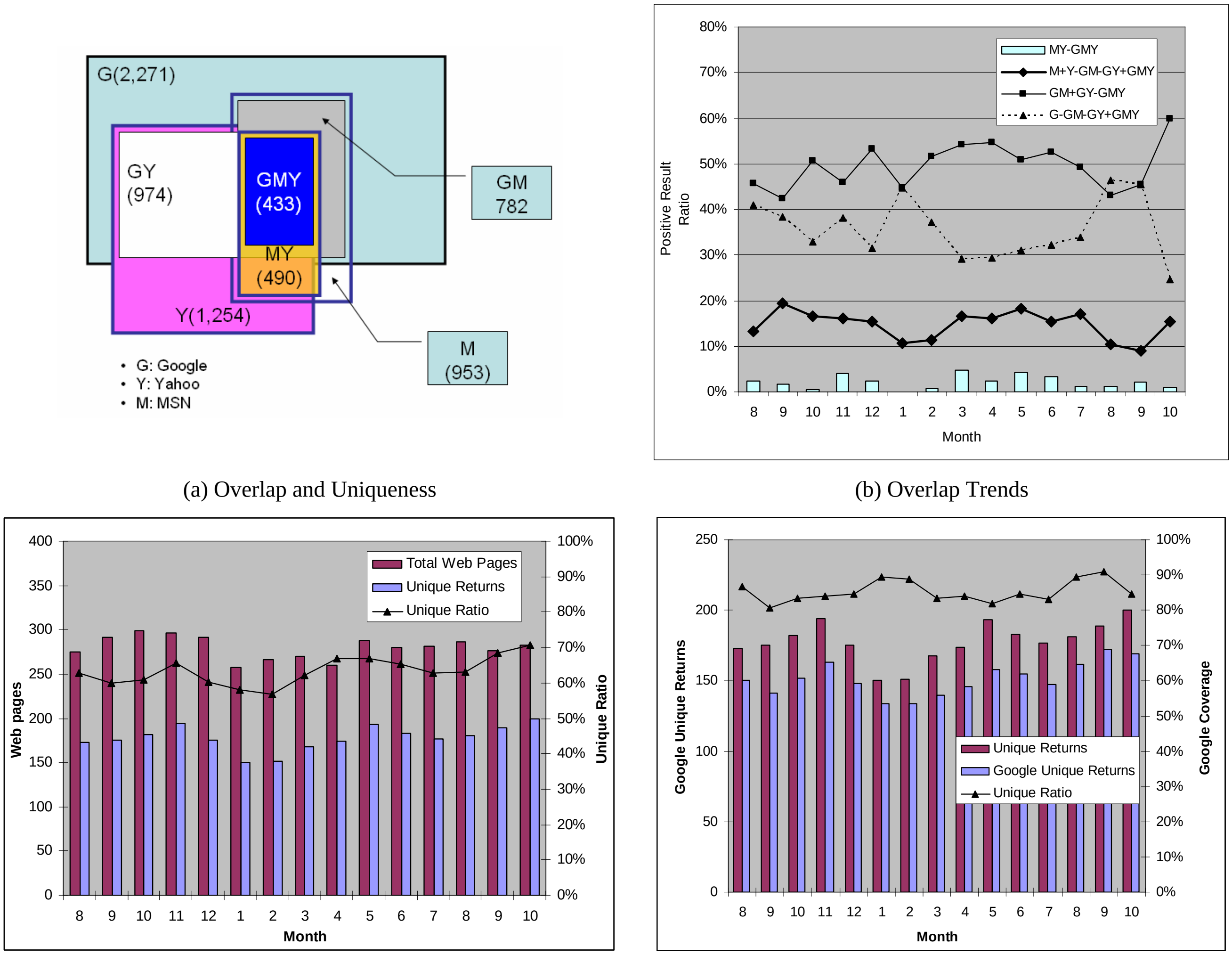
- Pure positive results for Google (P_g) = Google – (MSN ∩ Yahoo) + (Google ∩ MSN ∩ Yahoo) = (G-GM-GY-GYM)
- Pure positive results for MSN and Yahoo ($P_{m,y}$) = (MSN ∪ Yahoo) –(Google ∩ MSN) – (Google ∩ Yahoo) – (MSN ∩ Yahoo) + (Google ∩ MSN ∩ Yahoo) = (Y+M–GM–GY–MY+GMY)
- Pure overlap of Google and MSN and Yahoo(O_p) = (Google ∩ MSN) + (Google ∩ Yahoo) – (Google ∩ MSN ∩ Yahoo) = GM+GY-GMY

Ratios of these three sections against total positive results are illustrated in Figure 7 (a). The sum of these three ratios is 100% because P_g + $P_{m,y}$ + O_p equals the total unique positive results. This analysis suggests a meta-search engine would provide 10% to 20 % higher coverage than most efficient search engine.

(5) The trends of the unique positive returns are very similar to the monitored Web information page trends. The unique positive returns trends over the monitoring period are displayed in the Figure 7 (c), where the bar graphs represent the sampled Web information pages and the unique positive returns and the line graph represents the ratio of unique positive return to Web information pages. The trends of the unique positive returns are very similar to the monitored Web information page trends, except for recent months.

(6) The overlap and uniqueness data reflect the characteristics of the Web information source pages. Table 4 (b)

summarises overall overlap and uniqueness for the four different areas. The total unique positive return (TUPR) ratio for local government is greater than for federal government. We conjecture that crawlers give more priority to federal government Web pages than to local government, as federal government web pages cover larger domains and are probably revisited more often. The TUPR ratio for media release is greater than for homepages and here we hypothesise that since homepages tend to change their contents more slowly than media release pages, they are more likely to be visited by all the search engines. On the other hand Google's high return of media release pages suggests that Google might revisit web pages with more links more frequently. This is all speculation, but the one thing that is clear is that web crawling is a complex task controlled by a range of heuristics that vary with different search engines.



(a) Overlap and Uniqueness

(b) Overlap Trends

(c) Unique Positive Results Trends

(d) Google Coverage for Unique Results

Figure 7. Overlap and Uniqueness

6. DISCUSSION

6.1 Coverage: Missed Web Information Pages

The above coverage results show that the Web search engines missed from 46 % to 77 % of the Web information pages linked from the Web information source pages. We suggest there are three main reasons for this.

(1) **Crawler's coverage problem**: Crawlers may miss some Web information simply because they do not know the existence of Web information source pages. This problem is inevitable because of the passive information delivery of the internet. They do not become aware of a Web information source until the page is linked from a known Web information source pages or manually reported by the publishers.

(2) **Crawlers' revisit scheduling problem:** Ideally crawlers should revisit Web information source pages to get new Web information pages before they disappear from the source pages. The disappearance of new Web information pages depends on various factors such as frequency of publication, the number of URLs number that the Web information source page can reasonably contain, and the advent of specific events. Furthermore, deleting, moving, and updating published Web pages happens frequently [15, 16].

(3) **Web information source page duplication problem**: Duplication of Web information source pages also

contributes to reduced coverage. Web information can be posted by more than two Web information source pages. If crawlers have the capability to eliminate duplicate Web information from different Web information sources, they may be missed because our definition of a positive return requires an exact match of the URL given.

6.2 Overlap, Dominance and Meta-Search Engines

Meta-search engines try to integrate various search engines to provide better coverage. They are based on two assumptions – low overlap among search engines and low dominance of search engines. Our results do not show these sorts of patterns in the domains we have investigated.

After Ding and Marchionini[9] first identified aspects of the low overlap among the search results, similar results were reported by other researchers. Bharat and Broder [10] estimated the overlap among four commercial Web search engines as 1.4% and similarly Spink et. al [8] found the overlap of Ask Jeeves, Google, MSN, and Yahoo to be 1.1%. Our research results a 9.7% overlap between three search engines. We conjecture our results are different for the following reasons. Firstly, Bharat and Broder and Spink et. al used query keywords that were generated randomly or were generated by users, we use more complete and exact query keywords because we know the Web pages what we want to find. Secondly, the differences might be due to the number of evaluated search results. Spink et. al evaluated only the first pages in the list returned because they assumed users tend to look at the first pages. However, measured overlap between search engines should not depend on ranking methods.

Low dominance of any one search engine is another long standing assumption. Lawrence and Giles[11] reported any single Web search engines indexed no more than 16% of all Web sites and Selberg and Etzioni [18] suggested that no single search engine is likely to return more than 45% of relevant results. However, our result shows that Google is clearly dominant in these domains, because its coverage for unique positive results is 85%.

For domains like ours, high overlap and high dominance weaken the value of meta-search as an integrator of Web search engines. They are no doubt useful for different types of pages from the ones we investigated and as well our results do not provide any information on the value of meta-search engines as integrators for different ranking systems, the other problem for which they have been proposed.

6.3 Timeliness

As discussed in section 6.1, crawlers miss new Web information pages because of their inappropriate revisit scheduling. When a new Web information page is published, generally it takes some time before it is collected by crawlers and serviced by the search engines.

Prior research mainly focuses on the changing characteristics of the Web itself. For example, Duglis et al. [19] analysed a collection of HTTP responses from two companies, the Digital and AT&T, to evaluate the rate and nature of changes in Web resources. They found that many Web resources change frequently and that the frequency of access, time since last modification, and frequency of modification depend on content type and top-level domain, but not size. Ntoulas et al.[20] crawled all pages from 154 sites and their results show that 8% of the pages are replaced and 25% new links are created every week. Fetterly et al.[21] suggested that "the average degree of change varies widely across top-level domains, and that larger pages change more often and more severely than smaller ones".

Brewington and Cybenko [13] conducted an empirical study of timeliness, but did not consider an adaptive scheduling mechanism. More elaborate scheduling algorithms were suggested by [5], [22], and [23]. They viewed the Web monitoring scheduling problem as a delay minimization problem given a resource allocation policy. They used indirect published time (e.g., last modified date in HTTP header) for their scheduling algorithm. It would be more useful to establish publishing time using a Web monitor program. As part of our further research we will investigate adaptive scheduling algorithms and further investigate patterns relating Web information page publication and the Web crawler revisit behaviour.

Table 4 Overlap and Uniqueness by Areas

	Google	Yahoo	MSN	G&Y	G&M	Y&M	G&Y&M	TUPR
Federal Home	153	87	106	73	92	52	50	179[*](51.7%)[†]
Federal Media	1,316	930	700	722	575	380	333	1,602[*](54.4%)[†]
Local Home	258	135	115	88	91	44	38	323[*] (63.6%)[†]
Local Media	544	102	32	91	24	14	12	561[*](82.7%)[†]
Total	2,271	1,254	953	974	782	490	433	2,665 [*](59.5%)[†]

Note:

[*] Total Unique Positive Return (TUPR) are calculated by using the following formula
TUPR = G+Y+M–GM–GY–YM+GYM

where G, Y, M, GM, GY, YM, and GYM represent positive results from Google, Yahoo, MSN, and their overlapped positive returns of each domain.

[†] TUPR ratio is calculated using the following formula
TUOR ratio = TUPR / (G+Y+M)

where G,Y,M represents positive returns from each search engine of each area.

7. CONCLUSIONS

In this paper we studied coverage, overlap and dominance of three commercial search engines (Google, Yahoo, and MSN) using 15,770 Web information pages, which were collected from 260 Australian federal and local government Web pages for 15 months. We found that

(1) overall coverage of all three commercial search engines is 63.4% and individually they vary from 22.7% to 54.0%,

(2) overall overlap is 9.7%, which is large compared to other studies[8, 10]

(3) one search engine (Google) is dominant over other search engines, and covers 85% of all unique search returns.

We need to enhance coverage by employing dynamic scheduling strategy or use other Web information technologies such as Web monitoring and we need to reconsider the value of meta-search, because our results, especially (2) and (3), weaken the meta-search research assumption of the low coverage of each search engine and low dominance by any one search engine.

8. ACKNOWLEDGMENTS

This work is supported by the Asian Office of Aerospace Research and Development (AOARD) (Contract Number: FA5209-05-P-0253)

9. REFERENCES

[1] Duan, Z. and K. Gopalan, Push vs. Pull: Implications of Protocol Design on Controlling Unwanted Traffic. 2005, Florida State University.

[2] Martin-Flatin, J.P., Push vs. Pull in Web-Based Network Management. 1998, Switzerland: Lausanne.

[3] Cho, J., Crawling the Web: Discovery and Maintenace of Lagre-Scale Web Data, in Department of Computer Science. 2001, Stanford University.

[4] Tang, W., L. Liu, and C. Pu. WebCQ detecting and delivering Information changes on the Web. in Proc. Int. Conf. on Information and Knowledge Management (CIKM). 2000.

[5] Pandey, S., K. Dhamdhere, and C. Olston. WIC: A General-Purpose Algorithm for Monitoring Web Information Sources. in 30th VLDB Conference. 2004. Toronto, Canada.

[6] Douglis, F., et al., The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. World Wide Web, 1998. 1(1): p. 27-44.

[7] Boyapati, V., et al. ChangeDetector[tm]: a site-level monitoring tool for the WWW. in WWW 2002. 2002.

[8] Spink, A., et al., A study of results overlap and uniqueness among major web search engines. Information Processing and Management, 2006. 42(5): p. 1379 - 1391.

[9] Ding, W. and G. Marchionini. A Comparative Study of Web Search Service Performance. in Annual Conference of the American Society for Information Science. 1998.

[10] Bharat, K. and A. Broder. A technique for measuring the relative size and overlap of public Web search engines. in WWW7: The Seventh International World Wide Web Conference. 1998. Brisbane, Australia.

[11] Lawrence, S. and C.L. Giles, Searching the World Wide Web. Science, 1998. 280.

[12] Simon, H.A., Computers, Communications and the Public Interest, ed. M. Greenberger. 1971: The Johns Hopkins Press. 335.

[13] Brewington, B.E. and G. Cybenko, Keeping Up with the Changing Web. Computer, 2000. 33(5): p. 52-58.

[14] Koehler, W., Web page change and persistence - A four-year longitudinal study. Journal of the American Society for Information Science and Technology, 2001. 53(2): p. 162 - 171.

[15] Bar-Ilan, J. and B.C. Peritz, Evolution, continuity, and disappearance of documents on a specific topic on the web: a longitudinal study of "Informetrics". Journal of the American Society for Information Science and Technology, 2004. 55(11): p. 980 - 990.

[16] Fetterly, D., et al. A large-scale study of the evolution of web pages. in 12th international conference on World Wide Web. 2003. Budapest, Hungary.

[17] Park, S.S., S.K. Kim, and B.H. Kang. Web Information Management System: Personalization and Generalization. in the IADIS International Conference WWW/Internet 2003. 2003.

[18] Selberg, E. and O. Etzioni. On the Instability of Web Search Engines. in RIAO 2000. 2000.

[19] Douglis, F., A. Feldmann, and B.Krishnamurthy. Rate of change and other metrics: a live study of the world wide web. in the USENIX Symposium on Internet Technologies and Systems. 1997. Monterey, California.

[20] Ntoulas, A., J. Cho, and C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. in WWW 2004. 2004. New York, NY USA: ACM.

[21] Fetterly, D., et al. A large-scale study of the evolution of web pages. in 12th international conference on World Wide Web. 2003. Budapest, Hungary: Publisher ACM Press New York, NY, USA.

[22] Pandey, S., K. Ramamritham, and S. Chakrabarti. Monitoring the dynamic web to respond to continuous queries. in International World Wide Web Conference. 2003. Budapest, Hungary.

[23] Sia, K.C. and J. Cho, Efficient Monitoring Algorithm for Fast News Alert. 2005, UCLA Computer Science Department.