What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots

Masashi Toyoda
toyoda@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
kitsure@tkl.iis.u-tokyo.ac.jp

Institute of Industrial Science, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, JAPAN

ABSTRACT

Identifying and tracking new information on the Web is important in sociology, marketing, and survey research, since new trends might be apparent in the new information. Such changes can be observed by crawling the Web periodically. In practice, however, it is impossible to crawl the entire expanding Web repeatedly. This means that the novelty of a page remains unknown, even if that page did not exist in previous snapshots. In this paper, we propose a novelty measure for estimating the certainty that a newly crawled page appeared between the previous and current crawls. Using this novelty measure, new pages can be extracted from a series of unstable snapshots for further analysis and mining to identify new trends on the Web. We evaluated the precision, recall, and miss rate of the novelty measure using our Japanese web archive, and applied it to a Web archive search engine.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurements

Keywords

Link analysis, Web evolution, novelty, information retrieval

1. INTRODUCTION

The dramatic growth in the volume and diversity of the Web has enhanced the possibility of using the Web as a tool for sociology, marketing, and survey research. People in these disciplines, such as sociologists and market researchers, are interested in how the Web evolves over time based on events in the real and virtual worlds, and try to observe and track trends on their topics through the Web.

Identifying and tracking new information on the Web is important for such research, since new trends might be apparent in the new information. For example, we can determine when and how topical keywords emerge and evolve over time, by plotting occurrences of keywords in newly appeared pages. This technique can also be used to examine the penetration of brands or concepts on the Web.

Basically, the evolution of the Web is examined by periodic crawling. Using a time series of crawled Web snapshots, its evolution can be examined, such as changes in page content, increases in the number of pages on topics, and changes in the hyperlink structure. Recent research on the evolution of the Web is based on ideal crawls. For example, in [10], a fixed set of pages was crawled periodically to examine changes in these pages, and in [19], all of the pages comprising 154 popular web sites were crawled periodically to identify newly appeared pages on these sites.

In practice, however, such periodic crawling has inherent difficulties in tracking the evolution of the Web. Even with ideal crawls, some aspects of its evolution will be missed. If the set of pages is fixed, the appearance of new pages will be overlooked. If the set of sites is fixed, the appearance of new pages on these sites will be captured, but the appearance of new sites will be overlooked. To capture such evolution, it is necessary to crawl a massive amount of the Web during each period. However, this causes difficulty in keeping snapshots stable. Recently, it has become essentially impossible to crawl every web page on the Web. As mentioned in [11], the number of uncrawled pages overwhelms the number of crawled pages, even after crawling over a billion pages, and there are numerous dynamic URLs generated by databases. Moreover, the evolution of the structure of the Web changes the order of link traversal drastically. These facts make it impossible to fully crawl the entire growing Web, and to build stable web snapshots.

In this situation, can we tell whether newly crawled pages are really new? Even if a page is found that did not exist in previous snapshots, the novelty of the page is still uncertain; it might have been misidentified as new because:

- The crawler did not access that page in the previous crawl.
- The site containing the page was temporarily unavailable due to sever or network troubles during the previous crawl.
- The page was not linked to the starting points of the previous crawl.

The only time information available for pages is the Last-Modified time in its HTTP response. However, this guarantees only that the page is older than that time, and gives no indication of when the page first appeared.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

In this paper, we propose a novelty measure for estimating the certainty that a newly crawled page p appeared between a previous and current crawls. Basically, we consider that page p does not exist on the Web until p can be retrieved by search engines. This means that page p appears when it can first be crawled by following links from existing pages. Page p can therefore be considered novel when it is pointed to only by links that have appeared since the previous crawl. The novelty measure is discounted when links for which the creation time is unknown point to p. In addition, if only novel pages point to p, p may also be novel. This implies a recursive definition of the novelty measure, and it can be calculated in a similar way to PageRank [7], although the purpose of our ranking is quite different. The novelty measure also has an inverse relationship with the decay measure [2], which represents the decay of a page in terms of the probability that a random surfer reaches a dead page. This means that the novelty measure can be calculated in a similar way to the decay measure, except for inverted link direction. While both PageRank and the decay measure are calculated using a single snapshot, our novelty measure is calculated using temporal changes in multiple Web snapshots. We evaluate the precision and recall of our novelty measure using our Japanese web archive, and show that novel pages can be identified with reasonable precision and recall.

Using the novelty measure, novel pages can be extracted from a series of unstable snapshots for further analysis and mining to find novel trends on the Web. This study presents an application of the novelty measure to our archive search engine, displaying changes in the number of novel pages related to query keywords, and ranking search results according to their novelty.

The rest of this paper is organized as follows. Section 2 discusses related work. The notion of the novelty measure is described in Section 3. In Section 4, we present our experiments and evaluations. Section 5 shows an application of the novelty measure to our archive search engine. Finally, Section 6 concludes the paper.

2. RELATED WORK

Brewington and Cybenko [6], and Cho and Garcia-Molina [10] studied the frequency of changes and the lifetime of web pages. They estimated the frequency of web page modifications, and showed that the results were useful for web crawlers to determine the timing of further crawls. Fetterly *et al.* [12] extended these studies by crawling a set of 150 million pages once every week, and showed various statistics for changes in the pages. Since they focused on the changes in the pages, their results were based on periodic crawls of a specific set of pages; consequently, they could not detect the appearance of new pages. Ntoulas *et al.* [19] periodically crawled all the pages on popular 154 sites, and examined the volume of newly appeared information on these sites. However, such crawling is possible only for a small subset of the Web.

There has been much work on the Web graph structure and modeling its evolution, for example [9, 16, 5, 22, 17]. The macroscopic structure of the Web was studied in [9, 16], and a site (or server) level analysis of Web graph evolution was studied in [5]. In [22, 17], emerging cyber communities were extracted based on a graph evolution model.

Based on the structure of the Web, various link analysis methods have been proposed for information retrieval, such as the PageRank algorithm [7] by Brin and Page, and the HITS algorithm [14] by Kleinberg. These algorithms and their variants calculate the relevance and relationships of web pages. Recently, Bar-Yossef *et al.* [2] proposed a different measure of the page importance, the so-called decay measure, to reveal the death and decay of web pages. The decay is explained in terms of the probability that a random surfer stopping at a dead page. Its calculation process is similar to that of PageRank, but it can be calculated without computation for all pages. Our novelty measure is similar to the decay measure, but differs in using temporal changes in the Web graph, and in the direction of random surfing.

Various studies have attempted to reveal trends in the Web. The Internet Archive [25] once provided a full-text search engine called Recall [20] that had a keyword search future for 11 billion pages in its archive. Based on the search results, Recall provided a graph showing changes in the frequency of the search keyword over time. From that graph, it was possible to determine when keywords emerged or faded. Amitay *et al.* [1] proposed a method to detect trends using time-stamped links. The authors assumed that links were time-stamped with the Last-Modified time of their source pages. In response to given keywords, they extracted the top pages of the search result and pages that pointed to the top results. Then, they confirmed the trend from a histogram of time-stamped links for this set of pages. Our previous work [23, 24] examined the evolution of web communities. A web community is a set of web pages with a common interest in a topic. In [23], we proposed a method for extracting all web communities and their relationships from a single web archive. In [24], we extracted all the web communities from periodically crawled web snapshots, and visualized changes in these communities, such as growth and shrinkage. These studies assumed that newly crawled pages were novel, and were not concerned with whether the pages really were novel. Recently, tracking trends in blogspace has become a hot topic (e.g., [15, 13]), since blogs are good information sources for extracting individual reputations and opinions. In blogs, all of the articles have already been time-stamped using blog tools, which makes it easier to track the evolution of topics. However, blogspace is a rather small subset of the entire Web, and we are interested in the more global evolution of topics, not only in blogs but also in pages of governments, companies, newspapers, magazines, etc.

3. NOVELTY MEASURE

Let $t_k (1 < k < n)$ be the time when each snapshot was crawled. Let $W(t_k)$ be the snapshot at time t_k, which is a set of crawled pages. Let $G(t_k) = (V(t_k), E(t_k))$ be the Web graph of $W(t_k)$ where nodes in $V(t_k)$ represent web pages, and edges in $E(t_k)$ represent links. $V(t_k)$ includes pages outside $W(t_k)$, if they are pointed to by pages in $W(t_k)$. Outside pages in the previous snapshots at least show their existence in the past.

In the following, we first classify pages in $W(t_k)$ as old pages $O(t_k)$ and unidentified pages $U(t_k)$, where $W(t_k) = O(t_k) \cup U(t_k)$ and $O(t_k) \cap U(t_k) = \emptyset$. $O(t_k)$ is the set of pages that existed before t_{k-1}. $U(t_k)$ is the set of pages that were newly crawled at t_k, but their novelty remains unknown. Then, we define the novelty measure $\mathcal{N}(p)$ of a page p in $U(t_k)$. The score of $\mathcal{N}(p)$ is a number between 0 and 1. The score 1 represents the highest certainty that p appeared between t_{k-1} and t_k. The score 0 represents the

novelty of p as totally unknown; note that it does not mean that p is old. Finally, we describe the relationship between the novelty measure and the decay measure.

Old pages can be identified by checking previous snapshots, and by checking the Last-Modified time in the HTTP responses. A page is old when it exists in previous snapshots or if old pages point to it. Formally, page p is in $O(t_k)$ when:

- p is in $V(t_j)$ for existing $j \leq t_{k-1}$,
- The Last-Modified time of p is earlier than t_{k-1}, or
- p is pointed to by q with a Last-Modified time earlier than t_{k-1}.

Then, we define the novelty measure $\mathcal{N}(p)$ for each page p in $U(t_k)$ using the novelty of links pointing to p. $\mathcal{N}(p)$ represents the certainty that p appeared between t_{k-1} and t_k. Basically, we consider that page p does not exist on the Web until p can be retrieved by search engines. This means that page p appears when it can be crawled by following links from existing pages. Formally, page p appears at t_k when p is pointed to only by links that newly appeared between t_{k-1} and t_k.

The novelty of a link (q, p) can be identified by changes in the source page q as follows:

- The link (q, p) was created between t_{k-1} and t_k, if the source page q is in $O(t_k)$, and was crawled at both t_{k-1} and t_k (i.e. $q \in W(t_{k-1})$ and $q \in W(t_k)$). Since we know the contents of q at both t_{k-1} and t_k, and that there was no link from q to p at t_{k-1} from the definition of $U(t_k)$ (p is not in $V(t_j)(j < k)$). In the following, we use $L_2(t_k) \subseteq O(t_k)$ as the set of pages that was crawled at both t_{k-1} and t_k.
- If q is in $O(t_k)$ and does not follow the above condition, the novelty of the link (q, p) is uncertain because we cannot know when that link was created. This case can occur when q was crawled intermittently or q was outside $W(t_j)(j < k)$. This makes the novelty of p somewhat uncertain.
- If q is in $U(t_k)$, the novelty of link (q, p) depends on the novelty of q. That is, if q is identified as novel, (q, p) is also novel.

This implies a recursive definition of the novelty measure as follows:

$$\mathcal{N}(p) = (1-\delta)\frac{\sum_{(q,p)\in I(p)} n(q,p)}{|I(p)|}$$

$$n(q,p) = \begin{cases} 1 & q \in L_2(t_k) \\ 0 & q \in O(t_k)\backslash L_2(t_k) \\ \mathcal{N}(q) & q \in U(t_k) \end{cases} \quad (1)$$

where $I(p)$ is the set of links in $E(t_k)$ pointing to p. The parameter δ represents the probability that there were links to p before t_{k-1} outside the snapshots, and works as a damping factor that decreases the novelty measure as it propagates to other pages. It follows the intuition that the certainty of novelty decreases with the distance from novel pages.

We can rewrite this definition of the novelty measure without using the novelty of links. This refinement is important for calculating the novelty measure, since we do not need memory space to store the score for each link (usually, the number of links overwhelms the number of pages). We temporarily assign the constant novelty 1 to pages in $L_2(t_k)$, and 0 to other pages in $O(t_k)$. Then the definition of $\mathcal{N}(p)$ becomes the following:

$$\mathcal{N}(p) = \begin{cases} 1 & p \in L_2(t_k) \\ 0 & p \in O(t_k)\backslash L_2(t_k) \\ (1-\delta)\frac{\sum_{(q,p)\in I(p)} \mathcal{N}(q)}{|I(p)|} & otherwise. \end{cases} \quad (2)$$

Interestingly, from this refined definition, we can see the inverse relation between the novelty measure and the decay measure [2]. This means that the main difference is the direction of score propagation. The decay measure assigns the score 1 to dead pages, and propagates scores to backward pages with a damping factor. Conversely, the novelty measure assigns the score 1 to the sources of newly created links, and propagates scores to forward pages with a damping factor.

As with the decay measure, the novelty measure can be considered as the absorption probabilities in a random walk on the Web graph, in which the direction of each link is inverted, and the success and failure states are added. When a walk reaches a page in $L_2(t_k)$, it is absorbed in the success state at the next step (with probability 1). When a walk reaches a page in $O(t_k)\backslash L_2(t_k)$, it is absorbed in the failure state at the next step. In addition, from all pages in $U(t_k)$, walks are absorbed in the failure state with the probability δ. Therefore the novelty measure of page p is the absorption probability in the success state when starting from page p.

Time	Period	Crawled pages	Links
1999	Jul to Aug	17M	120M
2000	Jun to Aug	17M	112M
2001	Oct	40M	331M
2002	Feb	45M	375M
2003	Feb	66M	1058M
2003	Jul	97M	1589M
2004	Jan	81M	3452M
2004	May	96M	4505M

Table 1: Number of pages and links in each web snapshot.

Time	Jul 2003	Jan 2004	May 2004
$\|L_2(t_k)\|$	49M	61M	46M
$\|O(t_k)\backslash L_2(t_k)\|$	23M	14M	20M
$\|U(t_k)\|$	25M	6M	30M
$\|W(t_k)\|$	97M	81M	96M

Table 2: The number of old and unidentified pages in each snapshot.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

For the experiments, we used large-scale snapshots of our Japanese web archive, built during periodic crawls conducted between 1999 and 2004 (See Table 1). Basically, our crawler is based on breadth-first crawling [18]; except that it focuses on pages written in Japanese. Until 2002, we collected

Time(t_k)	Jul 2003		Jan 2004		May 2004	
Old ($t < t_{k-1}$)	299,591	(33%)	87,878	(24%)	402,365	(33%)
New ($t_{k-1} \leq t \leq t_k$)	593,317	(65%)	270,355	(74%)	776,360	(64%)
Future ($t_k < t$)	24,286	(2%)	7,679	(2%)	36,476	(3%)
Total	917,194	(100%)	365,912	(100%)	1,215,201	(100%)

Table 3: The number of unidentified URLs ($U(t_k)$) including time strings.

pages in the `.jp` domain. Beginning in 2003, we started to collect pages outside the `.jp` domain if they were written in Japanese. We used a web site as a unit when filtering non-Japanese pages. The crawler stopped collecting pages from a site, if it could not find any Japanese pages on the site within the first few pages. In our experiments, we used three snapshots obtained between July 2003 and May 2004 to examine the novelty measure, since there were no serious changes in the crawling strategy in that period.

From each snapshot, we built databases of URLs and links by extracting anchors from all of the pages in the snapshot. Our link database stores $G(t_k)$ for each snapshot. To check old URLs, we also built a database of the URL status of each snapshot, which stores whether each URL was crawled or was on the fringe of the Web graph, and also stores the time when the URL was crawled and when it was last modified. For efficient link analysis, each link database was implemented as a main-memory database that provides the out-links and in-links of a given URL. The implementation is similar to [4, 21], except that the same ID is used for each URL over time.

4.2 Old and Unidentified Pages

First, we show that there are a significant number of unidentified pages in each snapshot. Table 2 shows the number of old and unidentified pages in each snapshot. To calculate the number of old pages, we used all of the snapshots from 1999. In July 2003 and May 2004, about half of the pages were crawled during the last two crawls ($L_2(t_k)$). The percentage of unidentified pages was 25% and 30% in in July 2003 and May 2004, respectively. This means that these portions of snapshots were replaced by newly crawled pages.

Our snapshots are somewhat unstable. The number of crawled pages was rather small in Jan 2004; 75% of pages were in $L_2(t_k)$, and only 7% were unidentified. In this crawl, our crawler tended to collect old pages, and could not crawled new pages as well as in the other crawls.

4.3 Convergence of Calculation

The novelty measure is calculated in almost the same way as the PageRank [7] calculation. We computed the novelty scores of all unidentified pages ($U(t_k)$) in a batch process as follows:

1. Identify all old pages ($O(t_k)$), and assign initial scores to them according to definition (2) in Section 3. That is, if page p is in $L_2(t_k)$, the initial score of p is 1, and the score is 0 in other cases.

2. For all pages in $U(t_k)$, calculate the novelty measure according to definition (2). This calculation is repeated iteratively until the novelty of all pages converges.

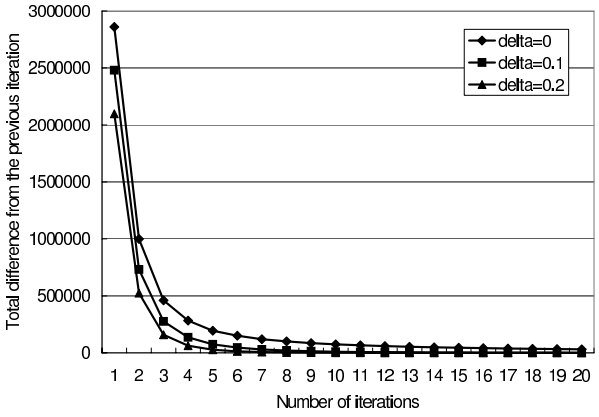

Figure 1: Convergence of the novelty measure calculation for different delta parameters (May 2004 snapshot).

Figure 1 shows the convergence of the novelty measure calculation with the May 2004 snapshot for δ values of 0.0, 0.1, and 0.2. The x-axis is the number of iterations, and the y-axis is the total difference in the novelty measure from the previous iteration. This shows that in practice 10 to 20 iterations are sufficient for any δ value. The convergence is rather slow when using $\delta = 0.0$, and at least 20 iterations are needed for convergence. This is because propagated scores are not dampened by iterations. Using 0.1 and 0.2, convergence becomes faster, and 10 iterations are sufficient for practical use. In the following, we use the novelty measure after 20 iterations.

4.4 Distribution of the Novelty Measure

In Figures 2 to 4, we show the distribution and cumulative curve of the novelty measure for pages in $U(t_k)$ calculated using different δ values. Each distribution has two peaks, one around 0.0 and the other at the maximum value, because the calculation of the novelty measure propagates 0 and 1 values of old pages with the damping factor δ.

The overall distribution does not change drastically with the δ value, while the maximum value of the novelty measure changes with the δ value. When δ increases, the maximum value of $\mathcal{N}(p)$ decreases by δ.

These distributions are sensitive to the fraction of $L_2(t_k)$ and $U(t_k)$. The peak around the maximum value rises as the fraction of $L_2(t_k)$ increases and that of $U(t_k)$ decreases (Jan 2004), since more 1 values are propagated to a smaller set of $U(t_k)$. The peak around the maximum value falls as the fraction of $L_2(t_k)$ decreases and that of $U(t_k)$ increases (July 2003, and May 2004).

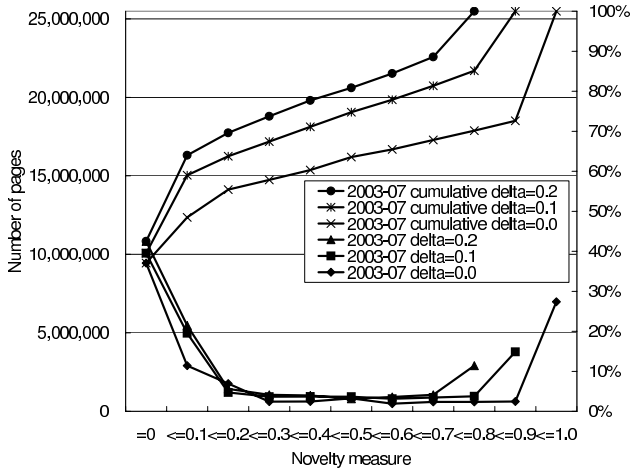

Figure 2: Distribution of the novelty measure for newly crawled pages in Jul 2003.

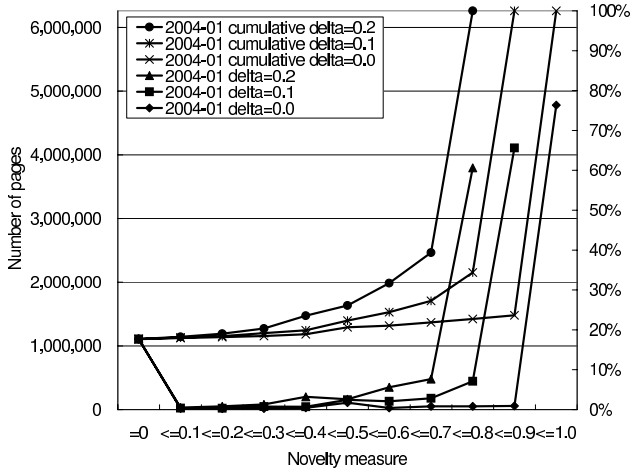

Figure 3: Distribution of the novelty measure for newly crawled pages in Jan 2004.

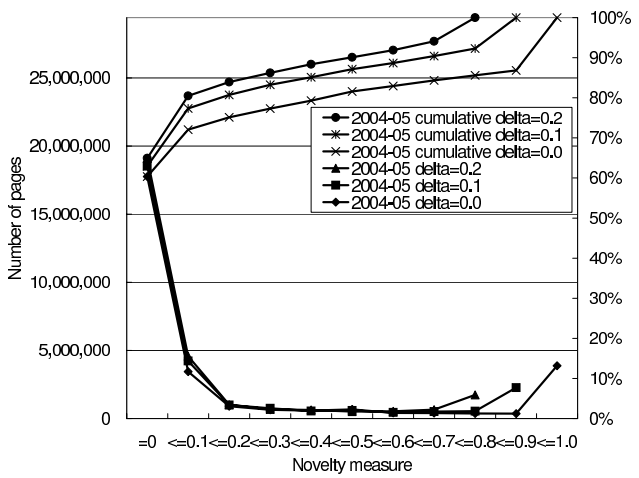

Figure 4: Distribution of the novelty measure for newly crawled pages in May 2004.

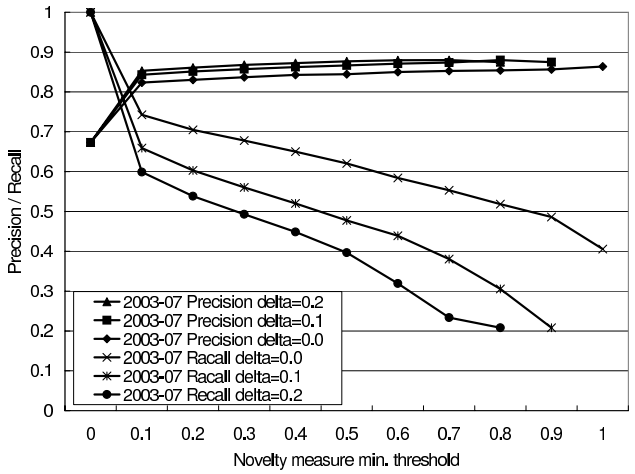

Figure 5: Precision and recall of the novelty measure in Jul 2003.

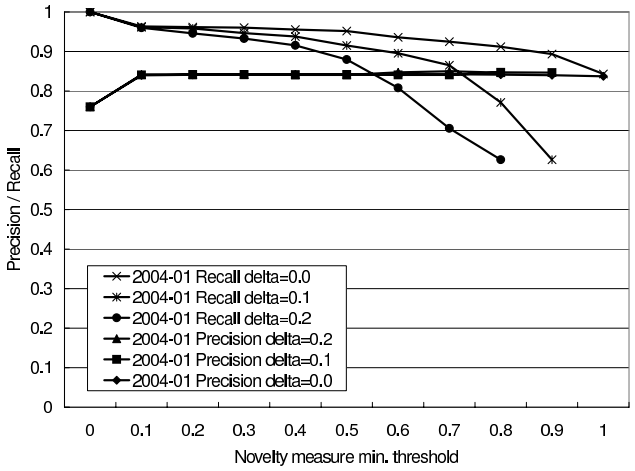

Figure 6: Precision and recall of the novelty measure in Jan 2004.

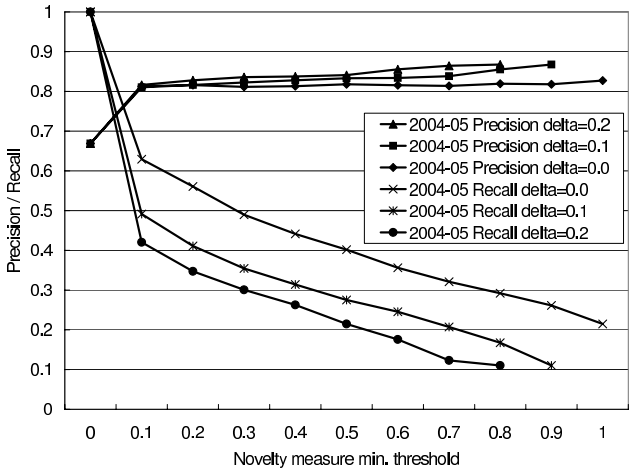

Figure 7: Precision and recall of the novelty measure in May 2005.

4.5 Precision and Recall

In practice, we judge a page to be novel if the novelty measure of the page exceeds some threshold value. In this experiment, we evaluated the precision and recall of this judgment by changing the minimum threshold value. The precision in this context is the ratio of the number of novel pages to the number of pages judged to be novel. The recall is the ratio of the number of pages judged to be novel to the number of all novel pages.

Since it is impossible to build accurate benchmarks, we perform a rough estimation using URLs that include time strings, such as `foo.com/2004/05/` and `bar.com/200405/`. We assume that such URLs appeared at their included times, and use them as answers. Table 3 shows the number of URLs in $U(t_k)$ including time strings. We extracted URLs that include a year (from 1993 to 2010) and a month in the following patterns:

- YYYYMM
- YYYY/MM
- YYYY–MM

Table 3 also shows the percentages of old URLs (with time strings before t_{k-1}), new URLs (between t_{k-1} and t_k), and future URLs (past t_k). We use the new and future URLs as novel URLs for measuring the precision and recall. If we judge all URLs to be novel, the precision values are 0.67, 0.76, and 0.67 for July 2003, Jan 2004, and May 2004, respectively. These precision values are baselines for evaluation.

This benchmark is partially incorrect, since some URLs might have appeared before or after their included times. However, the probability that a URL appeared before its time is small. The number of future URLs in Table 3 shows that this probability for each snapshot was only a few percent. We cannot know the probability that a URL appeared after its time, but it is also expected to be small.

Figures 5 to 7 shows the precision and recall for different δ values by changing the minimum threshold. The x-axis is the threshold value. We judge a page to be novel when the novelty measure of the page is greater or equal to the threshold. The y-axis shows the precision and recall.

In all snapshots, any positive threshold gives higher precision than the baseline. In Figures 5 to 7, the precision jumps from the baseline when the threshold becomes positive. Using a higher threshold value makes the precision slightly higher, but this increase does not exceed the first jump. Using positive δ values gives slightly better precision than using $\delta = 0.0$, but the difference between 0.1 and 0.2 is essentially zero.

The recall decreases drastically as the threshold increases in Figures 5 and 7, but the decrease is slower in Figure 6. This arises from the distribution of the novelty (See 2 to 4). In the July 2003 and May 2004 snapshots, the number of pages with low novelty scores exceeds the number of pages with high novelty scores. This makes the decrease of the recall faster. Using a positive δ value also decreases the recall, since it slows the novelty score propagation (See Section 3).

When using the novelty measure to extract novel pages, one has to select appropriate parameters. If higher precision is required, one should use a positive δ value, and a higher threshold value. If higher recall is required, one should use $\delta = 0$, and a small positive threshold value.

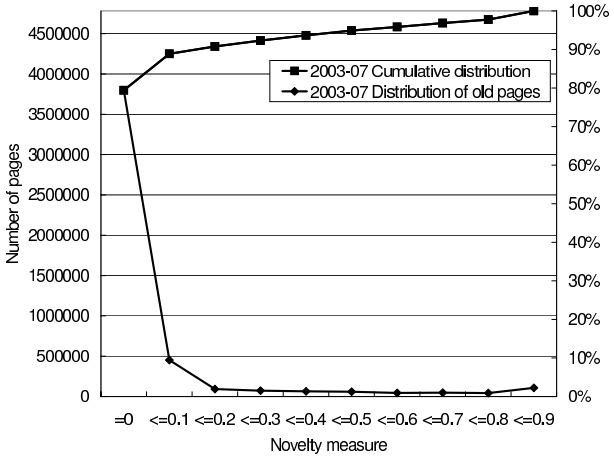

Figure 8: Distribution of the novelty measure for old and unidentified pages with the Last-Modified time before Feb 2003 in the Jun 2003 snapshot.

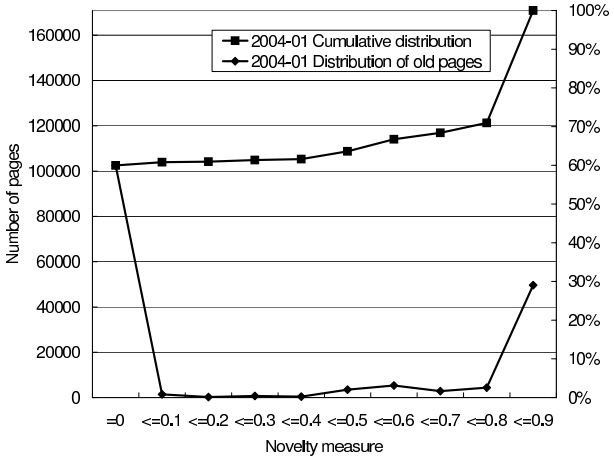

Figure 9: Distribution of the novelty measure for old and unidentified pages with the Last-Modified time before Jul 2003 in the Jan 2004 snapshot.

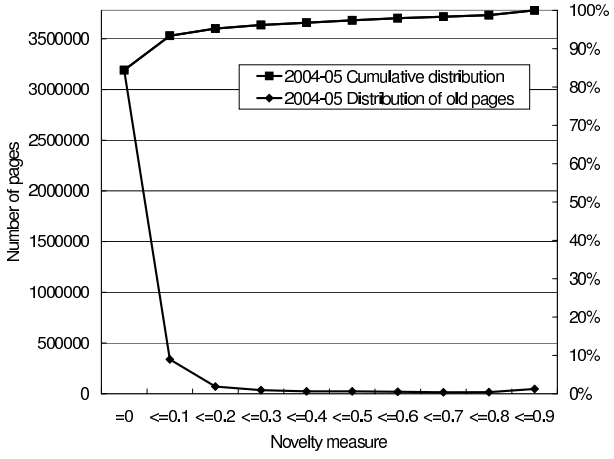

Figure 10: Distribution of the novelty measure for old and unidentified pages with the Last-Modified time before Jan 2004 in the May 2004 snapshot.

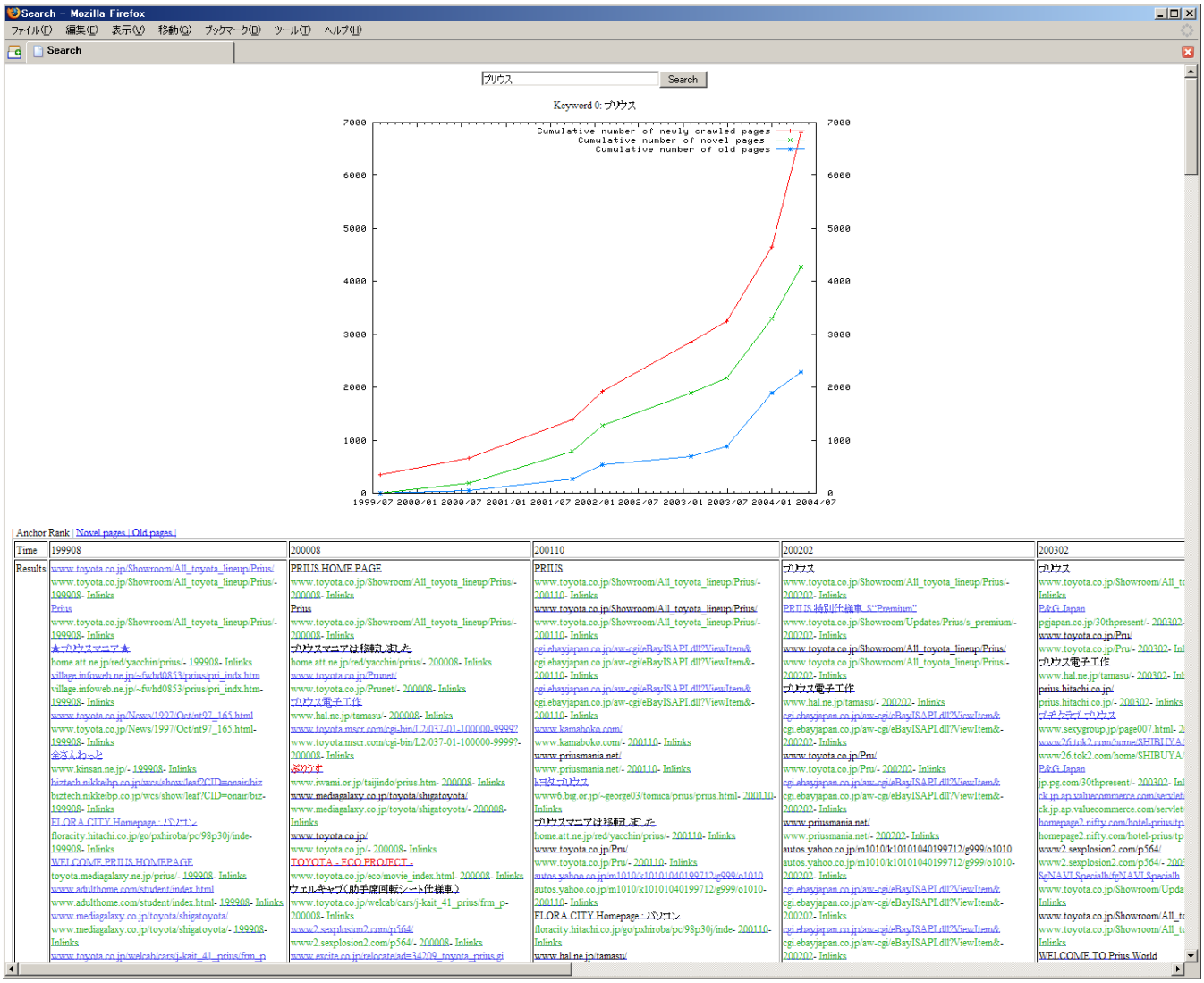

Figure 11: Screen snapshot of the search engine interface for our Web archive.

4.6 Miss Rate

Finally, we evaluate the miss rate of the novelty measure using the Last-Modified time of pages. We know that page p is older than t_{k-1}, if the Last-Modified time of p is before t_{k-1}. According to the Last-Modified time, we extract a set of old pages, and examine how many of the pages have been assigned positive novelty measures. The definition of the miss rate is the ratio of the number of pages judged to be novel to the number of all old pages.

Figures 8 to 10 shows the distribution of the novelty measure for old and unidentified pages in each snapshot. The novelty measure was calculated using 20 iterations with $\delta = 0.1$. In this calculation, we do not use the Last-Modified time for identifying old pages (See Section 3). Compared with the overall distributions (in Figures 2 to 4), these distributions are shifted largely to lower values. This shows that old pages tend to be assigned low novelty values.

We can see the miss rate from cumulative curves. In July 2003 and May 2004, there are 3.5 to 4.5 million pages that are old and unidentified. About 80% of these old pages are assigned 0 as the novelty value in both snapshots. If we judge the pages with positive scores to be novel, the miss rate is about 20%. Using 0.1 as the minimum threshold, the miss rate decreases to about 10%. In Jan 2004, the miss rate is about 40%, but the effect is limited, because there are only 0.17 million old and unidentified pages. This is only 3% of all unidentified pages.

5. APPLICATION

Currently, we are building a search engine for our Web archive that uses the novelty measure to display changes in the number of novel pages related to query keywords, and to rank search results. The engine searches documents according to given keywords in each snapshot. Figure 11 is a screen snapshot of the search engine. It includes a graph that shows cumulative curves of the number of novel pages over time. In keyword search results, we should consider three types of novel page:

- Old pages in which the keywords are newly appeared,
- Newly crawled pages that are judged to be novel using the novelty measure, and
- Newly crawled pages that are judged to be uncertain using the novelty measure. (Note that a low novelty score does not mean that the page is old.)

In the graph in Figure 11, the numbers of these three types of pages are stacked and accumulated over time.

The results of the keyword search are provided for each snapshot, as shown in the bottom of Figure 11. These results are normally sorted using a simple ranking method. They can also be sorted using the novelty measure in ascending or descending order. By sorting the results in descending order of the novelty measure, a user can extract novel pages for further analysis and mining to find novel trends.

Figure 11 shows search results using the keyword "Prius", which is the name of a hybrid vehicle made by Toyota. We can see that the number of novel pages jumped in Jan 2004, as a result of a full model change in Sep 2003. By sorting the results for Jan 2004 by the novelty measure, we can extract pages about the model change, such as news releases and blogs.

6. CONCLUDING REMARKS

In this paper, we proposed the notion of a novelty measure for identifying newly appeared pages in a series of unstable Web snapshots. By using the novelty measure, a set of novel pages can be extracted with reasonable precision and recall for further analysis and mining to identify novel trends on the Web. We also applied the novelty measure to our archive search engine, and showed an example of tracking novel pages on the Web.

The novelty measure can also be used for a kind of focused crawling that collects mainly novel pages. Such focused crawling is important to find emergent information according to the user's interest. Once the novelty measure has been calculated for the current snapshot, the focused crawler can use the distribution of the novelty measure to collect novel areas from the Web.

The novelty measure is available for pages that are generated dynamically only if they have unique URLs that do not change over time. Recently, such permanent URLs are supported in most web publishing tools, such as blogs and Wikis. Estimating the novelty of the deep Web [3] is another major research area, and is beyond the scope of this paper.

The novelty measure cannot discriminate copied, moved, and mirrored pages. Therefore, a page with old information might be judged to be novel if the page has newly appeared but its contents are old. Such pages should be preprocessed or post-processed using a mirror detection method, such as [8].

The snapshots used in our experiments are rather small subsets of the entire Web, and the crawling interval is quite long. We are interested in applying the novelty measure to more global snapshots, and in investigating how changing the snapshots affects the results. We are now crawling larger portions of the Web more frequently to try to observe finer-grained evolution.

7. ACKNOWLEDGMENTS

This work was partially supported by the *Comprehensive Development of e-Society Foundation Software* program of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

8. REFERENCES

[1] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Trend Detection through Temporal Link Analysis. *Journal of the American Society for Information Science and Technology*, 55(14):1270–1281, 2004.

[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic Transit Gloria Telae: Towards an Understading of the Web's Decay. In *Proceedings of the 13th International World Wide Web Conference*, pages 328–337, 2004.

[3] M. K. Bergman. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1), 2001.

[4] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The Connectivity Server: Fast Access to Linkage Information on the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 14–18, 1998.

[5] K. Bharat, B.-W. Chang, M. Henzinger, and M. Ruhl. Who Links to Whom: Mining Linkage between Web Sites. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 51–58, 2001.

[6] B. E. Brewington and G. Cybenko. How Dynamic is the Web? In *Proceedings of the 9th International World Wide Web Conference*, pages 257–276, 2000.

[7] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, 1998.

[8] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic Clustering of the Web. In *Proceedings of the 6th International World Wide Web Conference*, pages 391–404, 1997.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. In *Proceedings of the 9th International World Wide Web Conference*, pages 309–320, 2000.

[10] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB)*, pages 200–209, 2000.

[11] N. Eiron, K. S. McCurley, and A. Tomlin. Ranking the Web Frontier. In *Proceedings of the 13th International World Wide Web Conference*, pages 309–318, 2004.

[12] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A Large-Scale Study of the Evolution of Web Pages. In *Proceedings of the 12th International World Wide Web Conference*, pages 669–678, 2003.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion through Blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, pages 491–501, 2004.

[14] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the*

ACM-SIAM Symposium on Discrete Algorithms, pages 668–677, 1998.

[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the Bursty Evolution of Blogspace. In *Proceedings of the 12th International Conference on World Wide Web*, pages 568–576, 2003.

[16] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic Models for the Web Graph. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 57–65, 2000.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting Large-Scale Knowledge Bases from the Web. In *Proceedings of the 25th International Conference on Very Large Databases (VLDB)*, pages 639–650, 1999.

[18] M. Najork and J. L. Wiener. Breadth-First Crawling Yields High-Quality Pages. In *Proceedings of the 10th International World Wide Web Conference*, pages 114–118, 2001.

[19] A. Ntoulas, J. Cho, and C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In *Proceedings of the 13th International World Wide Web Conference*, pages 1–12, 2004.

[20] A. Patterson. CobWeb Search. http://ia00406.archive.org /cobwebsearch.ppt.

[21] K. Randall, R. Stata, R. Wickremesinghe, and J. Wiener. The Link Database: Fast Access to Graphs of the Web. *SRC Research Report 175*, Compaq Systems Research Center, 2001.

[22] S. R. Ravi Kumar, Prabhakar Raghavan and A. Tomkins. Trawling the Web for Emerging Cyber-Communities. In *Proceedings of the 8th International World Wide Web Conference*, pages 403–415, 1999.

[23] M. Toyoda and M. Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. In *Proceedings of the Twelfth Conference on Hypertext and Hypermedia (Hypertext 2001)*, pages 103–112, 2001.

[24] M. Toyoda and M. Kitsuregawa. Extracting Evolution of Web Communities from a Series of Web Archives. In *Proceedings of the Fourteenth Conference on Hypertext and Hypermedia (Hypertext 2003)*, pages 28–37, 2003.

[25] Wayback Machine, The Internet Archive. http://www.archive.org/.