Identify Temporal Websites Based on User Behavior Analysis

Yong Wang, Yiqun Liu,
Min Zhang, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China
wang-yong05@mails.thu.edu.cn

Liyun Ru
Sohu Inc. R&D center
Beijing, 100084, China
ruliyun@sohu-rd.com

Abstract

The web is growing at a rapid speed and it is almost impossible for a web crawler to download all new pages. Pages reporting breaking news should be stored into search engine index as soon as they are published, while others whose content is not time-related can be left for later crawls. We collected and analyzed into users' page-view data of 75,112,357 pages for 60 days. Using this data, we found that a large proportion of temporal pages are published by a small number of web sites providing news services, which should be crawled repeatedly with small intervals. Such temporal web sites of high freshness requirements can be identified by our algorithm based on user behavior analysis in page view data. 51.6% of all temporal pages can be picked up with a small overhead of untemporal pages. With this method, web crawlers can focus on these web sites and download pages from them with high priority.

1 Introduction

Many web users prefer accessing news reports from search engines. They type a few key words about a recent event and navigate to detailed reports about this event from the result list. Users will be frustrated if a search engine fails to perform such service and turn to other search engines to get access to news reports. In order to satisfy the users' needs, many search engines, including Google and Yahoo!, provide special channels for news retrieval and their web crawlers have to download newly appeared pages as soon as possible. However, the web is growing exponentially. The amount of new pages emerging every week is 8% of the whole web[Ntoulas et al., 2004]. It is almost impossible to download all novel pages in time.

Only a small proportion of novel pages are temporal. They report recent events and should be downloaded immediately, others which are untemporal can be downloaded later when it is convenient. So many search engines have different types of web crawlers to download the web with different policies. A common crawler checks updates of existing pages and crawls untemporal novel pages of all kinds of web sites with a relatively low frequency, usually once a month. Common crawlers are widely adopted by most search engines, but they are not suitable for news web sites which produce a great amount of pages every day. To news pages, there will be a large gap between their publication time and downloading time. Users can not get access to news pages in time. Thus another kind of crawler called instant crawler is developed. This crawler only focuses on temporal novel pages and checks updates of news web sites with much smaller intervals. Most newly-arrived content which is of high news value can be discovered by the instant crawler. Task distribution is shown in Figure 1.

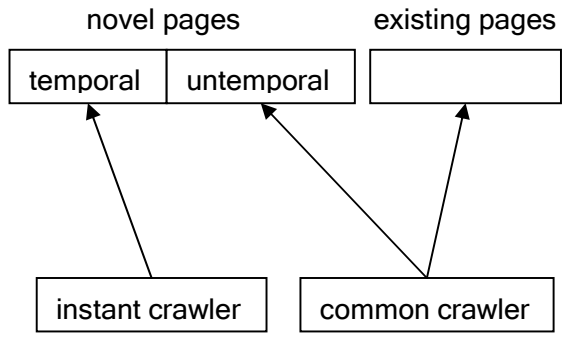

Figure 1. Job assigned to different crawlers

A relatively small set of web sites which provide news reporting services collectively generate many temporal web pages. These sites are valuable for instant crawlers and can be identified with web pages they previously generated. If a large proportion of web pages in a site are temporal, it is probable that pages published later from this the site will be temporal. Instant crawlers can focus on a list of such web sites.

Currently, the list of web sites for instant crawlers is usually generated manually, which is inevitably subjective and easily influenced by crawler administrators' preference. It includes many web sites which are actually untemporal. Also there are many mixed web sites which have both types of web pages. It is difficult for administrators to make accurate judgments about whether such sites should be included in the seed list. So instant crawlers have to spend precious and limited bandwidth to download untemporal pages while miss many temporal ones. What is more, this manually generated list is not sensitive to emerging and disappearing news sites.

In this paper, we propose a method to separate temporal pages from untemporal ones based on user behavior analysis in page-view data. Temporal web page identification is the prerequisite for temporal web site identification. A web site is temporal if most pages it publishes are temporal and most of its page-views are received from temporal pages. Then all web sites are ranked according to how temporal they are. Web sites ranked at a high position are included in the seed list for instant crawlers. Instant crawlers can focus on web sites in the list and only download pages from these web sites. Such a list covers a large proportion of temporal pages with only a small overhead of untemporal pages. The result is mined from web user behavior log, which reflects users' preference and avoids subjectivity of crawler administrators. Additionally, there are web sites associated with special events, such as Olympic Games. These web sites are temporal only when Olympic Games are being held. User behavior data can reflect the appearance and disappearance of temporal web sites.

An outline for the rest of the paper is as follows: Section 2 introduces earlier research in the evolution and discoverability of the web; Section 3 presents the user interest modal to describe web page lifetime from web users' perspective, then gives the definition of temporal web pages based on this model; Section 4 provides a method to generate a seed list for instant crawlers, and its result is also evaluated in the section; Section 5 discusses some alternatives in the experiment; Section 6 is the conclusion of this paper and suggests some possible directions in our future work.

2 Related Work

Earlier researchers performed intensive study on properties of images of the web graph[Barabasi and Albert, 1999; Broder et al., 2000; Kumar et al., 1999; Mitzenmacher, 2004]. Recently, researchers turned their attention to how the web evolves, including the rates of updates of existing pages[Brewington and Cybenko, 2000; Cho and Garcia-Molina, 2000; Fetterly et al., 2004; Pitkow and Pirolli, 1997] and the rates of new page emergence [Brewington and Cybenko, 2000]. They sent out a crawler to download web pages periodically, compared local images of the web and found characteristics of web page lifetime. Some researchers studied the frequency of web page update, predicted the lifetime of web pages and recrawl the already downloaded pages when necessary to keep the local repository fresh so that users are less bothered by stale information. They assumed that pages are modified or deleted randomly and independently with a fixed rate over time, so lifetimes of web pages are independent and identically distributed and a sequence of modifications and deletions can be modeled by a Poisson process. Other researchers focused on the discoverability of new pages[Dasgupta et al., 2007]. They tried to discover as many new pages as possible at the cost of only recrawling a few known pages.

But the web is growing explosively. It is impossible to download all the new pages. A crawler faces a frontier of the web, which is consisted of a set of discovered but not

downloaded URLs (see Figure 2). The crawler has to make a decision about which URLs should be downloaded first, which URLs should be downloaded later and which URLs are not worthy downloading at all. Thus there is some work in ordering the frontier for a crawl according to the predicted quality of the unknown pages[Cho, Garcia-Molina and Page, 1998; Eiron et al., 2004]. They predicted quality of pages which have not been downloaded yet based on the link structure of the web.

This job is similar with ours. We also make an order of the frontier, in the perspective of freshness requirements, not in the perspective of page quality. Freshness requirements differ from pages to pages. Temporal web pages whose freshness requirement timescale is minute, hour or day are assigned to the instant crawler with high priority. Other pages of lower freshness requirements can be crawled later.

This study is conducted with user behavior data instead of link structure. The link structure of the web is controlled by web site administrators. It reflects the preference of web site administrators, not that of the web users. Although in many cases, the two kinds of preference are alike, they are not identical. User behavior data reveals the real needs of web users. What is more, link structure can be easily misled by spammers. But spammers can do little to influence user behavior data contributed by mass web users.

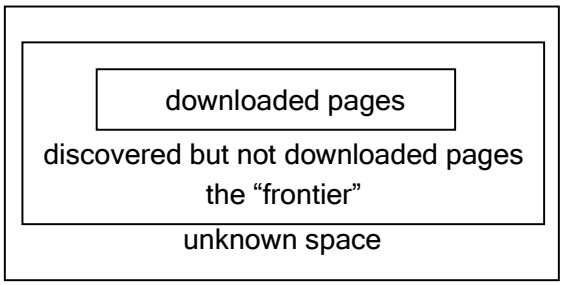

Figure 2. Web from a crawler's perspective

3 Definition of Temporal Web Page Based on Web Page Lifetime Model

3.1 Web page lifetime model

A page is born when it is published on a server, and it dies when it is deleted. But from users' view, its lifetime should not be defined by whether it is stored on a server but by whether it is accessed by users, because a web page is useful only when it can provide information for users. To web users, its life really starts at the "activation day" when the first user visits it. The page begins to be dormant on its "dormancy day" when users no longer visit it any more. After that, whether it is stored on the server does not make many differences. So the valid lifetime of a web page is the period between its activation day and its dormancy day, a subinterval of the period when it is accessible. Users' access is the only indication that the page is alive.

The state of a web page could be recorded with two values: alive and dead[Dhyani, 2002; Fetterly et al., 2003; Cho and Garcia-molina, 2003]. Its state during its valid lifetime is more complex and its liveness could be described with a continuous value: user interest. The number of page views it receives differs every day. It is more active when it is accessed by many users and it is not so active when it is accessed by fewer users. The amount of page views it receives reflects how many users are interested in it.

User interest in a web page is an amount indicating to what extent web users as a whole are interested in the page. A user visits a web page because he/she is interested in its content. The amount of page views a page receives is determined by how much user interest it can attracts. User interest in a page is a continuous variable evolving over time. User interest increases if more and more users get to know the page, and it decreases if the content is no longer fresh to users and the page becomes obsolete. User interest in a page whose content is not time related typically does not fluctuate greatly over time.

Web page lifetime could be described with user interest model, and then temporal web pages can be separated from untemporal ones according to different characteristics of their lifetime.

3.2 Definition of temporal web pages

There are two types of new pages: temporal pages and untemporal ones. Temporal pages are those reporting recent events. Users are interested in a temporal page and visit it only during a few hours or a few days after it is published. For example, a page reporting the president election result is temporal. Untemporal pages are the pages whose content is not related with recent events. There are always users visiting such pages. For example, a page introducing swimming skills is untemporal. The two kinds of new pages should be treated with different policies. The instant crawler has to download temporal pages as soon as they are discovered because users are interested in them

only in a short time span after they are born. Temporal pages are about new events and cannot be replaced by earlier pages. If the instant crawler fails to download temporal pages in time, search engine users cannot get the latest information because there are no earlier pages reporting the event which has just happened. One week after the event, even if temporal pages are downloaded, they are no longer attractive to users, just like a piece of old newspaper. In contrast, untemporal pages are not of exigencies. There is no need to download them immediately after they are published. Even if they are not downloaded in time, users can still be satisfied by other existing pages with similar content, since untemporal pages concern with problems which have already existed for a long time and have been discussed in many pages. It does not make many differences to download them early or a month later. So untemporal pages can be left to common crawlers to be downloaded later.

4 Temporal Web Sites Identification Algorithm

A seed list for an instant crawler contains temporal web sites. There are three steps to generate the seed list: search user's interest curves to describe web page lifetime; identify temporal web pages based on user interest curves; identify temporal web sites according to the proportion of temporal pages in each site.

4.1 Search User's Interest Curves

Generally speaking, few users know a newly born web page and pay attention to it. Later, more and more users get to know it, become interested in it and visit it. As time goes by, some pages become outdated and attract less user attention, while other pages never suffer from obsolescence. Users' interest in them is relatively constant. So the typical trend of user interest evolution is to increase at first then decrease in the shape of a rainbow, or to keep static. It is true that user interest in some pages experiences multi-climaxes. But it is very unlikely that those climaxes appear in our observing window of two months. Since we are studying short term web page lifetime, we do not consider user interest with multi-climaxes. The curve $y = f(x)$ that is used to describe the evolution of user interest should satisfy 4 conditions below(assuming the page is activated at time 0):

1) its field of definition is $[0, +\infty)$
2) $f(0) = 0$
3) $f(x) \geq 0$ in its field of definition
4) it has only one maximum

The probability density function (PDF) of logarithmic normal distribution is one of the functions that satisfy the conditions, so a modified edition of it, which will be addressed later, is used to describe the evolution of user interest during whole web page lifetime.

Anonymous user access log for consecutive 60 days is collected by a proxy server. Multiple requests to a single page in one day are merged as one request to avoid automatically generated large numbers of requests by spammers. Daily page view data of 75,112,357 pages from November 13th 2006 to January 11th, 2007 is recorded. Pages whose total page views during the 60 days are less than 60 (one page view each day on average) are filtered out because of lack of reliability, leaving 975,151 reliable ones. In order to retrieve user interest curves, we build a coordinate first, where the x-axis denotes time and y-axis denotes the number of page views. Given the daily page view data of a page, there are a sequence of discrete dots in the coordination (x_i, y_i), $i = 1, 2, ..., 60$, where $x_i=i$ is the *i*th day, y_i is the number of page views on the *i*th day. After that, the dots can be fitted with the formula

$$f(x) = A \times \varphi_{ln}(x) = A \times \frac{1}{\sqrt{2\pi}\sigma(x-b)} \times e^{-\frac{(\ln(x-b)-\mu)^2}{2\sigma^2}}$$

where A, b, μ, σ are parameters and $\varphi_{ln}(x)$ is the probability density function of logarithmic normal distribution. Given a page p and its page view history (x_i, y_i) $(i = 1, 2, ..., 60)$, the four parameters can be determined and the user interest curve can be defined as $y = f(x)$. One of the retrieved user interest curves is shown in Figure 3.

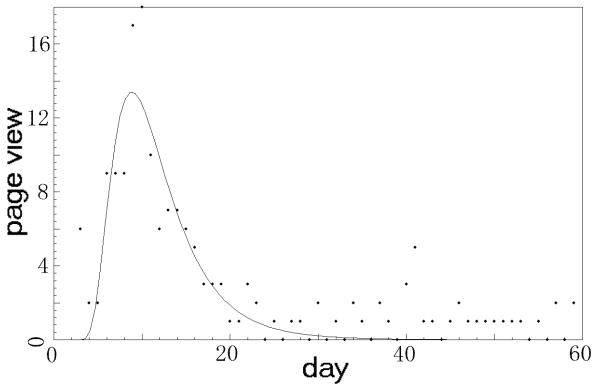

Figure 3. A User Interest Curve

4.2 Identifying Temporal Web Pages

$\varphi_{ln}(x)$ is the probability density function of a random variable. The integral of $\varphi_{ln}(x)$ in its field of definition is 1. The total user interest in a web page accumulated during its whole lifetime is

$$\int_b^{+\infty} f(x)\,dx = A \times \int_b^{+\infty} \varphi_{ln}(x)\,dx = A$$

The parameter A of a popular web page is larger than that of an unpopular one. In order to avoid discriminating popular pages and unpopular ones, parameter A for all pages is set to 1, so the area of the region enclosed by user interest curve and x-axis is 1. After this normalization, each page receives one unit user interest during their whole lifetime.

Parameter b indicates the birth time of a page. We do not care about the absolute birth time of a page, so parameter b for all pages are set to 0, which means all pages are activated at time 0.

The other two parameters σ and μ do not change, so the shape of user interest curves is reserved.

After the parameter adjusting, the user interest curve is redefined as

$$y = \varphi_{\ln}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \times e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

This simpler definition of user interest curve is used in the rest of this paper.

Let

$$\Phi(x) = \int_0^x \varphi(t)\,dt$$

be the cumulative density function of logarithmic normal distribution. Given the user interest curve of page p, Φ(x) is the amount of user interest accumulated x days after its birth (see the grey area in Figure 4).

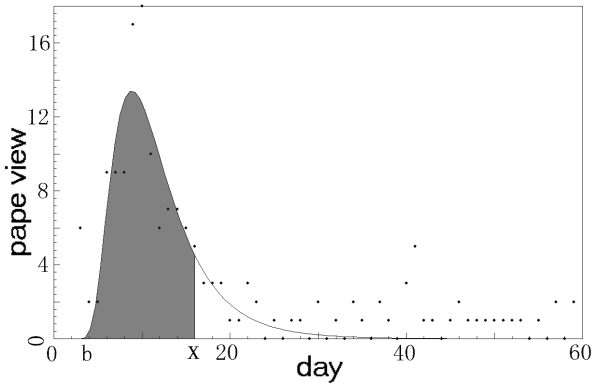

Figure 4. Accumulated user interest

A temporal web page accumulates most of its user interest during the first few days after its birth. Given a specific x, the larger Φ(x) is, the more temporal the page is, because it can accumulate more user interest during the time span. news.sohu.com is a major portal web site providing news services. Most of its pages are news reportings, which are temporal. There are 6,464 web pages from news.sohu.com and their user interest curves are retrieved. Figure 5 shows the distribution of Φ(1) of these pages. As is shown in Figure 5, on the first day of their birth, most pages have accumulated more than 80% of its total user interest of their whole lifetime. So the proportion of user interest accumulated during the beginning period of web page lifetime is a useful feature to identify temporal web pages.

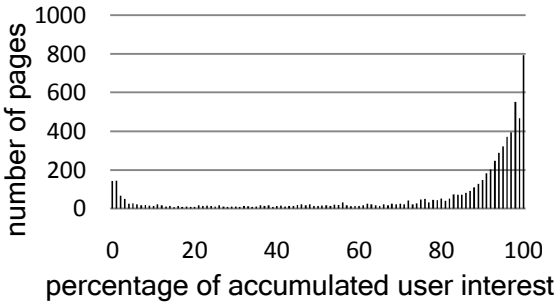

Figure 5. Distribution of accumulated user interest on the first day of birth

In order to discern temporal pages from untemporal ones, two parameters should be determined: n and q. n is the integrating range, q is the integral quantity and is also the grey area in Figure 4. Given a web page p, it is temporal if the proportion of user interest accumulated during the first n days of its lifetime is more than q (denoted in the inequality Φ(n) > q), vice versa. focus.cn is a web site about real estate. It publishes both temporal pages (such as those reporting price fluctuation information) and untemporal ones (such as those providing house decoration suggestions). We annotate 3,040 web pages in the web site of focus.cn manually, of which 2,337 are labeled temporal, 703 are labeled untemporal. After parameter adjusting, n is set to 3 and q is set to 0.7 in order to achieve the best performance that the maximized hit(the number of correct classification) is 2,829, miss(the number of temporal pages which are classified as untemporal) is 141, false alarm(the number of untemporal pages which are classified as temporal) is 70. It means that a web page will be classified as temporal if in the first three days after its birth, it can accumulate more than 70% of the total user interest it attracts during its whole lifetime.

After the classification, 135,939 web pages are labeled temporal and the other 839,212 pages are labeled untemporal.

4.3 Identifying Temporal Web Sites

A web site has many pages. There are hardly any web sites that publish temporal pages or untemporal pages exclusively. Instead, an actual web site usually contains both temporal pages and untemporal ones. For example, a web site about automobiles publishes temporal pages reporting that a new style of cars appears on the market, and it also publishes untemporal pages about how to take good care of cars. In order to classify web sites with mixed types of pages, we present definitions of temporal web sites.

From web sites administrators' view, a web site is temporal if most of its pages are temporal. So if the proportion of temporal pages of a web site in all its pages is large enough, the web site will be classified as temporal. According to this definition, the type of a web site can be controlled by its administrator. If he/she wants to make the web site temporal, he/she can publish more temporal pages. But how are these pages received by web users? Even most pages in a web site are temporal, if users pay little attention to them and are attracted mainly by untemporal ones, this web site cannot be classified as temporal. So a temporal web site should also be defined from web users' view.

From web users' view, a web site is temporal if most of its page views are received from temporal pages. Given a web site which contains both temporal pages and untemporal ones, if users are more interested in its temporal pages, the site is more likely to be classified as temporal.

Both of the two definitions above make sense. So a web site has two scores about how temporal it is based on the two definitions. The two scores are calculated in the following formulas

$$\mathrm{Score}_1(s) = \frac{\text{the number of temporal pages in s}}{\text{the number of total pages in s}}$$

$$\mathrm{Score}_2(s) = \frac{\sum_{tp \in \text{temporal pages in s}} \text{the number of page views of tp}}{\sum_{p \in \text{pages in s}} \text{the number of page views of p}}$$

where *s* is a web site. Score_1 is the proportion of temporal web pages in all pages from the web site. Score_2 is the proportion of page views received from temporal web pages in all page views to the site. Then the two scores are combined with different weights into the final score for the web site.

$$\mathrm{Score}(s) = \alpha \times \mathrm{Score}_1(s) + \beta \times \mathrm{Score}_2(s)$$

Web sites are ranked according to the final score in descending order. Search engines can pick the web sites ranked at high positions in the list as seeds for instant crawlers. They can pick as many seeds as their instant crawler is capable to monitor. In our experiment, we choose the top 100 web sites in the ranked list as temporal sites. Since there are 135,939 temporal pages and 839,212 untemporal ones in the data set, precision is defined as the proportion of temporal pages of the top 100 web sites in all pages of those sites, and recall is defined as the proportion of temporal pages of the top 100 web sites in all temporal ones in the data set. The ranked list is evaluated with the traditional IR evaluation criterion: F-Measure [Baeza-Yates and Ribeiro-Neto, 1999], which is calculated as

$$\mathrm{F-Measure} = \frac{2\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Parameter α and β are adjusted to improve F-Measure. When the ratio of α and β is 3:2, the maximized F-Measure is achieved at 0.615, where there are 70,110 temporal pages and 21,886 untemporal ones in the top 100 web sites in the ranked list.

4.4 Evaluation of the Temporal Web Site List

Human annotated results are always considered optimal in general experiment result evaluation in information retrieval. However, in our task, human annotator cannot make a perfect seed list for instant crawlers, because it is very difficult to decide whether a web site containing appropriate amount of temporal pages and untemporal ones. In contrast, the method we propose make decision not only by the proportion of temporal pages in a site, but also by how well each kind of pages are received based on the amount of page views they get. So the seed list generated from user behavior data can outperform that generated by humans.

Sohu Inc. has a manually generated list containing 100 seed web sites for its instant crawler. This list is evaluated with the method above. The 100 web sites in the list cover 59,113 temporal pages and 49,124 untemporal ones. The performance of automatic generated seed list for instant crawlers using our method is compared with that of manually generated list as the base line. The result is shown in Table 1.

Compared with the base line, the top 100 web sites in our seed list contain 18.6% more temporal pages than those in the manually generated list. The total burden of the instant crawler is also reduced by

15.0% since it downloads 16,241 less pages.

	Base line	Our method
Temporal Pages	59,113	70,110
Total Pages	108,237	91,996
Precision	54.6%	76.2%
Recall	43.5%	51.6%
F-Measure	0.484	0.615

Table 1. Evaluation of the two seed list for instant crawlers

5 Discussion

5.1 Advantages of using user interest curves

There are three advantages of using user interest curves instead of raw page-view data. First, the number of page views is determined by the amount of user interest a page receives, but they do not strictly equal. Page view data is affected by many random factors, such as whether it is weekday or weekend. These random factors are called "noise" in general. Such noise can influence the number of page views, but it is not the determinant factor. The number of page views is centered on the amount of user interest and fluctuate around it, because page view data is a combination of user interest and the noise. User interest curves are less bothered by such noise since the noise is effectively eliminated after data fitting. Second, although the observing window is two months wide, which is wide enough to cover lifetime of most temporal web pages, there are still many temporal ones whose lifetime is across the observing window boundaries. Such fragmented page view records will bring in mistakes in identifying temporal web pages. But if most part of the lifetime of a web page lies in the observing window, the data fitting process is able to estimate the absent page-view data and make up the missing part of user interest curve of its whole lifetime. So the effects brought by cross-boundary web pages can be reduced. Third, user interest curve is continuous and can be integrated to show the accumulated user interest to the page in a period of time.

5.2 Effects of using different parameter values

In our experiment, we used a single threshold n and q (see Section 5) and a page is classified as a temporal one if the user interest it accumulates during n days after its birth is greater than q. But web users receive different types of news at different speed. We notice that financial, entertainment, political and military news gets through users rapidly. These kinds of news become obsolescent to users quickly, usually only a few minutes or hours after they are published. So web pages reporting such news are ephemeral and they can draw users' attention only in a short period after their birth. In contrast, it takes much more time for users to know other kinds of news. For example, a web page reporting a volcano eruption far away from users may not be so attractive and has to spend much more time to accumulate the specific proportion of user interest. So maybe it is necessary to give different thresholds for different types of news.

In our experiment, we choose values of n and p in order to get the maximized hit (see Section 5). Some web crawlers may have abundant network bandwidth and want lower miss. Other crawlers whose network bandwidth is very limited are untolerant with false alarm. So the result of temporal page classification can be evaluated by linear combination with different weights

Performance = A × hit - B × miss - C × false alarm

Values of A, B and C can be determined according to the capacity of s crawlers.

Whether a web site is temporal is determined by the proportion of its temporal pages and the proportion of its page views received from temporal pages. The two proportions are combined with different weights α and β in order to get maximized F-Measure (see Section 6). However, to some extent, the measure of page-view proportion is misleading, because a hot event which receives a great deal of user attention is usually reported by several news agencies. It is of little value to download redundant reports from different web sites although they get many page views. Page-view data discriminates against pages reporting events which receive little attention. Most of these pages cannot be replaced by others because they are usually the only page reporting such events. Whether these pages can be correctly retrieved influence user experience greatly. Users often judge a search engine by whether the pages receiving low attention can be recalled. So the temporal page proportion should be assigned with additional weight to avoid such bias.

6 Conclusion and Future Work

The web is growing rapidly. It is impossible to download all new pages. Web crawlers have to make a decision about which pages should be downloaded with high priority. Previous

researchers made decisions according to page quality and suggested downloading pages of high quality first. They ignored the fact that temporal pages should be downloaded first. Otherwise, they will become outdated soon. It is better to download these temporal pages immediately in the perspective of freshness requirement.

Only a few web sites collectively publish a large proportion of temporal pages. In this paper, an algorithm is introduced to score each web site about how temporal it is based on page-view data which records user behavior. Web sites scored high are judged as temporal. An instant crawler can focus on temporal sites only. It can download more temporal pages and less untemporal ones in order to improve its efficiency.

Temporal web site identification can be done in finer granularity. There are several possible directions. Firstly, many web site administrators prefer distributing temporal web pages and untemporal ones in different folder. For example, pages stored under "/news/" are more likely to be temporal. Secondly, dynamic URLs (URLs that contain the character "?" and pairs of parameter and value) generated from the same web page, which are treated as different pages in the current work, are very likely to share the same timeliness. For example, if "/a.asp?p=1" is a temporal page, it is probable that "/a.asp?p=2" is temporal. In the future, we plan to study timeliness of web sites at folder level instead of site level.

References

Albert-László Barabási, Réka Albert. 1999. *Emergence of Scaling in Random Networks*, Science, Vol. 286. no. 5439, Pages 509 - 512.

Alexandros Ntoulas, Junghoo Cho, Christopher Olston. 2004. *What's new on the Web? The evolution of the Web from a search engine perspective.* Proceedings of the 13th conference on World Wide Web, Pages 1 - 12.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins and Janet Wiener. 2000. *Graph Structure in the Web*. Computer Networks, Volume 33, Issues 1-6, Pages 309-320.

Brian E. Brewington and George Cybenko. 2000. *How Dynamic is the Web?* Computer Networks, Volume 33, Issues 1-6, Pages 257-276.

Dennis Fetterly, Mark Manasse, Marc Najork and Janet Wiener. 2003. A Large-Scale Study of the Evolution of Web Pages. Proceedings of WWW03, Pages 669-678.

Devanshu Dhyani, Wee Keong NG, and Sourav S. Bhowmick. 2002. A Survey of Web Metrics. ACM Computing Surveys, Volume 34, Issue 4, Pages 469 – 503.

James Pitkow and Peter Pirolli. 1997. *Life, Death, and Lawfulness on the Electronic Frontier*. Proceedings of the SIGCHI conference on Human factors in computing systems, Pages 383–390, 1997.

Junghoo Cho , Hector Garcia-Molina and Lawrence Page. 1998. *Efficient Crawling Through URL Ordering*. Computer Networks, Volume 30, Number 1, Pages 161-172(12).

Junghoo Cho and Hector Garcia-Molina. 2000. *The evolution of the web and implications for an incremental crawler*. In Proc. 26th VLDB, Pages 200–209.

Junghoo Cho and Hector Garcia-Molina. 2003. *Effective Page Refresh Policies for Web Crawlers*. ACM Transactions on Database Systems, Volume 28 , Issue 4, Pages 390 – 426.

Michael Mitzenmacher. 2004. *A Brief History of Lognormal and Power Law Distributions*. Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing.

Nadav Eiron, Kevin S. McCurley and John A. Tomlin. 2004. *Ranking the Web Frontier*. Proceedings of the 13th international conference on World Wide Web, pages 309–318.

Ravi Kumar , Prabhakar Raghavan , Sridhar Rajagopalan and Andrew Tomkins. 1999. *Trawling the Web for Emerging Cyber-communities*. Proceeding of the eighth international conference on World Wide Web, Pages 1481-1493.

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.