Estimating the Change of Web Pages*

Sung Jin Kim¹ and Sang Ho Lee²

¹ Department of Computer Science, University of California, Los Angeles, USA
sjkim@cs.ucla.edu

² School of Computing, Soongsil University, Seoul, Korea
shlee@comp.ssu.ac.kr

Abstract. This paper presents the estimation methods computing the probabilities of how many times web pages are downloaded and modified, respectively, in the future crawls. The methods can make web database administrators avoid unnecessarily requesting undownloadable and unmodified web pages in a page group. We postulate that the change behavior of web pages is strongly related to the past change behavior. We gather the change histories of approximately three million web pages at two-day intervals for 100 days, and estimate the future change behavior of those pages. Our estimation, which was evaluated by actual change behavior of the pages, worked well.

Keywords: web page change estimation and web database administration.

1 Introduction

Many web applications, such as search engines, proxy servers, and so on, usually create web databases (collections of web pages) that enable users to retrieve web pages internally. As web pages on the web change actively, it is important that administrators update web databases effectively. To help administrators establish update policies, a number of researchers [1, 2, 3, 4, 5, 6, 7, 8] investigated the change behavior of web pages.

A good change estimation model certainly can help administrators manage the databases effectively. [2, 3] described the estimation method, based on a Poisson or memoryless process model (which assumes that changes of web pages are independent of the past changes). [1] also proposed the estimation method, using the change histories of web pages and the Poisson model together. [5] did not follow the Poisson process's change assumption, but did use only the change histories of web pages to make the databases keep the minimum number of obsolete pages. The previous researches for the estimation have directed their concern on the effective management of the entire databases and the contents modification of web pages.

In this paper, we describe simple estimation methods that can help administrators manage the set of web pages, which is a subset of the databases. In practice, administrators can divide web databases into a number of page groups in their own

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD). KRF-2006-214-D00136.

way such as locality, domain, subject, and so on. Our estimation can help administrators manage each of the groups with different update policies on the basis of the aggregated change histories of all the pages in the corresponding group. Our main assumption is that the change behavior of web pages is strongly related to the past one. By the way, the previous studies simply ignored the fact that the downloadable states of web pages can change irregularly. However, we take into account not only the modification issue but also the download issue of web pages. The goal of our estimation is to help administrators avoid unnecessarily requesting undownloadable or unmodified web pages, in the situation where they have to manage the databases composed of many page groups, and where they feel that requests to undownloadable and unmodified pages are wasteful under a limited network bandwidth and a time.

We present the two estimation formulae: $P_{(Y=a, N=b)DR_{Y=c, N=d}}$ and $P_{(Y=a, N=b)DR_{Y=c, N=d}}$. The former denotes the probability that a page that was downloaded a times for $(a+b)$ download attempts will be downloaded c times for $(c+d)$ download attempts. The latter denotes the probability that a page is modified a times for $(a+b)$ successful downloads will be modified c times for $(c+d)$ successful downloads. We also define the two change metrics: the download rates and the modification rates. The download rate represents how consistently a web page is downloaded successfully. The modification rate represents how frequently a web page is modified. When the estimation formulae are used for a web page in a group, the distributions of download rates and modification rates of all the web pages in the group are used for the probability computation.

For the evaluation of our estimation methods, we first divide approximately 3 million web pages detected from 34,000 Korean sites into the two groups, named the famous sites and the random sites. We obtain the distributions of download rates and modification rates for each of the two groups, after monitoring those pages at two-day intervals for 100 days. We estimate the numbers of downloads and modifications of the selected URLs in the future five crawls, using the numbers of successful downloads and modifications in the past five crawls. Finally, the estimation was evaluated with the actual change behavior of the estimated pages. The evaluation results show our estimations work well.

The paper is organized as follows. Besides this introduction, we define the metrics for representing the change behavior of web pages in section 2. We gather the change histories of web pages in section 3. Section 4 presents our estimation methods to predict the change behavior of the web pages. Section 5 contains the closing remarks.

2 Metrics for Representing the Change of Web Pages

In this section, we describe the metrics for measuring the change of web pages in terms of download and modification issues. Fig. 1 represents a simple crawling example, where there are 16 crawls. A hyphen (“-”) means that there is no download attempt since the URL is not available at that point. A black circle (“•”) means that we fail to download the corresponding web page. The contents of a downloaded page is denoted as a circled character (here, ㉠, ㉡, etc.). For example, on the fourth crawl where three pages corresponding to URLs A, C, and D are detected, we download two pages whose contents are ㉠ and ㉡, respectively.

URL	Crawl number															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
B	-	-	-	-	□	●	□	-	●	-	-	-	-	-	-	-
C	-	-	□	□	-	●	□	□	●	-	●	●	-	-	-	-
D	●	□	-	●	□	□	□	□	-	□	●	□	□	-	-	□

Fig. 1. Crawling example

Definition 1. The first detection number and the last detection number are crawl numbers on which a URL (or a web page) is detected first and last, respectively. The detection rate of a page is defined as “the number of detections / (the last detection number – the first detection number + 1)”

For example, the page for URL B is detected on the 5th, 6th, 7th, and 9th crawls. The detection rate of the page is $(4/(9-5+1)) = 0.8$. Because the URLs with low detection rates are not informative enough to be studied, we should analyze the change behavior of web pages whose detection rates are higher than a given threshold.

Definition 2. The download rate of a page is defined as “the number of successful downloads / the number of download requests”. The download recall of a page is defined as “the number of download requests / the total number of crawls”. The download rate represents how consistently the corresponding web page has been downloaded. We accept the download rate of a page whose download recall is higher than a given threshold.

For example, the page for URL C is detected eight times, and the page is downloaded on the 3rd, 4th, 7th, and 8th crawls. The download rate is $(4/8) = 0.5$. Since there are eight attempts to download the page in 16 crawls, the download recall C is $(8/16) = 0.5$.

Definition 3. Suppose that a page of URL u is downloaded on the i^{th} crawl, and that there is at least one download success with URL u prior to the i^{th} crawl. The current content of the page on the i^{th} crawl is defined as the content of a page that is downloaded on the i^{th} crawl. The previous content of the page on the i^{th} crawl is defined as the content of the page that is successfully downloaded latest prior to the i^{th} crawl. If the current and previous contents of the page on the i^{th} crawl are different from each other, the page of URL u is defined to be modified on the i^{th} crawl.

Definition 4. The modification rate of a page is defined as “the number of modifications / (the number of successful downloads – 1)”. The modification recall of a page is defined as “(the number of successful downloads – 1) / (the total number of crawls – 1)”. The modification rate represents how frequently a web page changes in terms of the page contents. For the computation of the modification rate, the page should be downloaded at least twice. We accept the modification rate of a page whose modification recall is higher than a given threshold.

For example, the current content of the page for URL D on the 10th crawl is □ and the previous content is ⑤. The page is modified on the 10th crawl. On the 12th crawl, the current and previous contents are of the same as ⑥. We say that the page is not modified on the 12th crawl. The page is downloaded nine times, which implies that eight comparisons take place. The page is modified twice on the 7th and 10th crawls. The modification rate and modification recall are $(2/(9-1)) = 0.25$ and $(8/15) = 0.53$, respectively.

3 Gathering the Change Histories of Web Pages

In this section, we obtain the distribution of download rates and the modification rates for our estimation. We monitored the 34,000 Korean sites at two-day intervals for 100 days (from January to March, 2004). Our robot [9] detected approximately 1.8 million URLs on each crawl and accumulated three million URLs after all 50 crawls. We regard each of those URLs as representing an individual web page. We grouped the monitored sites into the set of 4,000 famous sites and the set of 30,000 random sites in our own way. We gathered two kinds of change history sets: one for pages in the famous sites and the other for pages in the random sites. Because we visited a site every other day, we do not know if a page changes multiple times between our visiting intervals.

Table 1 shows the distribution of the download rates (DR) for the accumulated URLs satisfying the condition that the download recall is at least 0.2 and the detection rate is 0.9 or more. Table 2 shows the distribution of the modification rates (MR) of the pages in which the modification recalls were at least 0.2 and the detection rates were at least 0.9.

Table 1. Distribution of download rates

Famous sites				Random sites			
DR	Percent of URLs	DR	Percent of URLs	DR	Percent of URLs	DR	Percent of URLs
0	22.01%	0.50 ~ 0.59	0.07%	0	15.92%	0.50 ~ 0.59	0.03%
0.01 ~ 0.09	0.06%	0.60 ~ 0.69	0.07%	0.01 ~ 0.09	0.20%	0.60 ~ 0.69	0.08%
0.10 ~ 0.19	0.04%	0.70 ~ 0.79	0.09%	0.10 ~ 0.19	0.02%	0.70 ~ 0.79	0.07%
0.20 ~ 0.29	0.06%	0.80 ~ 0.89	0.33%	0.20 ~ 0.29	0.03%	0.80 ~ 0.89	0.35%
0.30 ~ 0.39	0.04%	0.90 ~ 0.99	9.53%	0.30 ~ 0.39	0.28%	0.90 ~ 0.99	7.34%
0.40 ~ 0.49	0.05%	1	67.66%	0.40 ~ 0.49	0.03%	1	75.65%
Total: 100.00%				Total: 100.00%			

A number of contents comparison methods, such as the byte-wise comparison method, the document shingling method [6], the TF-IDF cosine distance methods [10], and so on, can be used to measure change of a web page. The byte-wise comparison method, which compares the contents of web pages character by character, is the strictest one. Therefore, even though any trivial changes take place on a web page, the method regards the page as being modified. We obtained

modification rates by means of the byte-wise comparison method. Our experiment shows the maximum level of modification rates. If having used some other comparison method, we could obtain lower modification rates than those we had.

Table 2. Distribution of modification rates

Famous sites				Random sites			
MR	Percent of URLs	MR	Percent of URLs	MR	Percent of URLs	MR	Percent of URLs
0	63.85%	0.50 ~ 0.59	1.04%	0	66.07%	0.50 ~ 0.59	0.40%
0.01 ~ 0.09	15.16%	0.60 ~ 0.69	0.49%	0.01 ~ 0.09	9.37%	0.60 ~ 0.69	0.17%
0.10 ~ 0.19	4.38%	0.70 ~ 0.79	0.86%	0.10 ~ 0.19	1.99%	0.70 ~ 0.79	0.19%
0.20 ~ 0.29	2.59%	0.80 ~ 0.89	0.98%	0.20 ~ 0.29	0.92%	0.80 ~ 0.89	0.19%
0.30 ~ 0.39	1.94%	0.90 ~ 0.99	1.28%	0.30 ~ 0.39	0.33%	0.90 ~ 0.99	0.44%
0.40 ~ 0.49	0.98%	1	6.46%	0.40 ~ 0.49	0.26%	1	3.23%
Total: 100.00%				Total: 100.00%			

4 Estimating the Change of Web Pages

In this section, we present two estimation formulae computing the probabilities of how many times web pages in a group will be downloaded and modified in the future crawls, using the distributions of the download rates and the modification rates of all the web pages in the group. Our estimation postulates the following. First, the future change behavior of web pages is strongly related to the past change behavior of the web page. Second, history of downloadable states and history of modification are independent (even though that a web page is modified at a point means that the page is downloaded successfully at that point). Third, we hold sufficiently long period of the change histories. Possible estimation period of a page cannot exceed to the observation period of the page. Fourth, we hold appropriately recent period of the change histories. The change behavior of a web page is related equally to all the past change behaviors we hold. Fifth, we estimate the future change behavior of web pages at the same intervals as the observation.

Let $P(DR(x))$ of a URL denote the probability that a download rate value of the URL is x , where x is rounded to two decimals. Using Table 1, we can compute $P(DR(x))$ easily. For instance, $P(DR(0))$, the probability that the download rate of a page is 0, is 22.01%, and $P(DR(1))$ is 67.66%. In Table 1, we assume that instances are distributed uniformly in each range. Hence, we compute $P(DR(x))$ by dividing the percent corresponding to x by 9 when $0.01 \leq x \leq 0.09$, or 10 when $0.10 \leq x \leq 0.99$. For instance, $P(DR(0.55))$ is $0.07\%/10 = 0.007\%$.

Let $P_{(Y=a, N=b)DR_{Y=c, N=d}}$ of a URL denote the probability that a page that was downloaded a times for $(a+b)$ download attempts will be downloaded c times for $(c+d)$ download attempts. Before computing the probability, we have two preconditions. First, the sum of a , b , c , and d should be less than or equal to the total number of crawls (50 in this paper). Second, the intervals of $(a+b)$ and $(c+d)$

download attempts are the same as those of the monitoring interval (2 days in this paper). The probability is computed by the following formula:

$$P_{(Y=a, N=b)DR_{Y=c, N=d}} = \frac{P\left(DR\left(\frac{a+c}{a+b+c+d}\right)\right)}{\sum_{i=0}^{c+d} P\left(DR\left(\frac{a+i}{a+b+c+d}\right)\right)}$$

As an example, let us compute $P_{(Y=0, N=2)DR_{Y=0, N=3}}$, which is the probability that we will get three consecutive download failures when we have already failed to download the page twice in a row.

$$\begin{aligned} P_{(Y=0, N=2)DR_{Y=0, N=3}} &= \frac{P(DR(0))}{P\left(DR\left(\frac{0}{5}\right)\right) + P\left(DR\left(\frac{1}{5}\right)\right) + P\left(DR\left(\frac{2}{5}\right)\right) + P\left(DR\left(\frac{3}{5}\right)\right)} \\ &= \frac{22.01}{22.01 + 0.006 + 0.005 + 0.007} = 99.9\% \end{aligned}$$

Now we intuitively explain the estimation formula. As a simple case, suppose that we have failed to download a page three times in a row. On the next crawl, if we download the page successfully, the download rate of the page becomes 0.25. If we fail to download the page again, the download rate of the URL becomes 0. Then, we estimate the probability that the page is downloaded successfully to be ' $P(DR(0.25)) / (P(DR(0)) + P(DR(0.25)))$ '. The estimation formula above generalizes this idea.

Estimating modifications of a page is similar with estimating successful downloads. We use a table that shows the distribution of modification rates of web pages (see Table 2). Let $P(MR(x))$ of a page denote the probability that a modification rate value of the page is x . $P(MR(x))$ is computed similarly to $P(DR(x))$. Let $P_{(Y=a, N=b)MR_{Y=c, N=d}}$ of a page denote the probability that a page is modified a times for $(a+b)$ successful downloads will be modified c times for $(c+d)$ successful downloads. The probability is computed by the following formula:

$$P_{(Y=a, N=b)MR_{Y=c, N=d}} = \frac{P\left(MR\left(\frac{a+c}{a+b+c+d}\right)\right)}{\sum_{i=0}^{c+d} P\left(MR\left(\frac{a+i}{a+b+c+d}\right)\right)}$$

For the evaluation of $P_{(Y=a, N=b)DR_{Y=c, N=d}}$, we selected 581,608 pages from the famous sites and 838,035 pages from the random sites. The selected pages had been detected five times on the 46th to 50th crawl. With the number of successful downloads in the past five crawls, we estimate the number of successful downloads in the next five crawls. There are six possible cases such that a page will be downloaded 0, 1, 2, 3, 4, or 5 times, i.e., $P_{(Y=a, N=(5-a)DR_{Y=0, N=5})}$, $P_{(Y=a, N=(5-a)DR_{Y=1, N=4})}$, ..., and $P_{(Y=a, N=(5-a)DR_{Y=5, N=0})}$, respectively, where a is the number of successful downloads. And then, we actually have attempted to download the selected pages five times for 10 days at two-day intervals.

Fig. 2. Evaluation result of $P(Y=a, N=b | DR_{Y=c, N=d})$

Fig. 2 shows the evaluation results. The left bars represent the estimated numbers of successfully downloaded pages. The right bars represent how many pages were actually downloaded 0, 1, 2, 3, 4, or 5 times. In Fig. 2(a), the differences between the two bars in each case are 324, 86, 39, 64, 175, and 92, respectively. In Fig. 2(b), the differences are 129, 55, 35, 120, 732, and 813, respectively. Given the 581,608 pages in the famous sites and 694,583 pages in the random sites, the total number of the URLs that were incorrectly estimated is trivial (less than 0.07% and 0.11%).

Fig. 3. Evaluation result of $P(Y=a, N=b | MR_{Y=c, N=d})$

For the evaluation of $P(Y=a, N=b | MR_{Y=c, N=d})$, we selected 483,603 pages and 766,328 pages, which had been downloaded web pages without any failures in the 44th to 50th crawl. We estimated the number of modifications of the web pages in the next five crawls. Fig. 3 shows the result of the evaluation result. Both the bars represent the estimated and the real numbers of modifications. In Fig. 3(a), the differences between the two bars in each case are 16,360, 22,057, 250, 1,347, 194, and 6,988 (the sum is 47,196 pages), respectively. In Fig. 3(b), the differences are 26,759, 24,659, 5,039, 191, 3,630, and 6,760, respectively. The modifications for 23,598 pages (4.8%) and 33,518 URLs (4.4%) were incorrectly estimated. Although the proportion of incorrectly estimated URLs is higher than the first evaluation, we believe that the

correctly estimated portion is acceptable enough (95.2% of the pages in the famous sites and 95.6% of the pages in the random sites) to avoid unnecessary requests for web pages that have not been changed.

5 Closing Remarks

We presented estimation methods computing the probabilities of how many times a page will be downloaded and modified on the future crawls. Our estimation postulates that the future change behavior of a web page is strongly related to the past changes. After observing the change histories of approximately three million web pages at two-day intervals for 100 days, we estimated the change behavior of those pages. The estimation was evaluated with the actual change behavior. As the results of download estimation, only 0.07% and 0.11% of the selected pages in the two page groups are incorrectly estimated. As the results of the modification estimation, 4.8% and 4.2% of the selected pages in the groups were incorrectly estimated. We believe that our estimation methods certainly help administrators manage their databases composed of a number of groups and avoid unnecessarily requesting undownloadable or unmodified web pages.

References

1. Brewington, B., Cybenko, G.: How Dynamic is the Web? the 9th World Wide Web Conference (2000) 257-276
2. Cho, J., Garcia-Molina, H.: The Evolution of the Web and Implications for an Incremental Crawler. the 26th VLDB Conference (2000) 200-209
3. Cho, J., Garcia-Molina, H. (2003): Effective Page Refresh Policies for Web crawlers. *ACM Transactions on Database Systems*. **28(4)** (2003) 390-426
4. Dougllis, F., Feldmann, A., Krishnamurthy, B.: Rate of Change and Other Metrics: a Live Study of the World Wide Web. the 1st USENIX Symposium on Internetworking Technologies and System (1997) 147-158
5. Edwards, G., McCurley, K., Tomlin, J.: Adaptive Model from Optimizing Performance of an Incremental Web Crawler. the 10th World Wide Web Conference (2001) 106-113
6. Fetterly, D., Manasse, M., Najork, M., Wiener, J. L.: A large-scale study of the evolution of web pages. the 12th World Wide Web conference (2003) 669-678
7. Ntoulas, A., Cho, J., Olston, C.: What's New on the Web? The Evolution of the Web from a Search Engine Perspective. the 13th World Wide Web Conference (2004) 1-12
8. Toyoda, M., Kitsuregawa, M.: What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots. the 15th World Wide Web Conference (2006) 233-241
9. Kim, S.J., Lee, S.H.: Implementation of a Web Robot and Statistics on the Korean Web. *Springer-Verlag Lecture Notes in Computer Science* **2713** (2003) 341-350
10. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)
11. Dhyani, D., Ng, W.K., Bhowmick S.S.: A Survey of Web Metrics. *ACM Computing Survey*. **34(4)** (2002) 469-503
12. Huberman, B.A.. The Laws of the Web: Patterns in the Ecology of Information. MIT Press (2001)