Crawling the Infinite Web: Five Levels are Enough

Ricardo Baeza-Yates and Carlos Castillo

Center for Web Research, DCC
Universidad de Chile
{rbaeza,ccastillo}@dcc.uchile.cl

Abstract. A large amount of publicly available Web pages are generated dynamically upon request, and contain links to other dynamically generated pages. This usually produces Web sites which can create arbitrarily many pages. In this article, several probabilistic models for browsing "infinite" Web sites are proposed and studied. We use these models to estimate how deep a crawler must go to download a significant portion of the Web site content that is actually visited. The proposed models are validated against real data on page views in several Web sites, showing that, in both theory and practice, a crawler needs to download just a few levels, no more than 3 to 5 "clicks" away from the start page, to reach 90% of the pages that users actually visit.

1 Introduction

Most studies about the web refer only to the "publicly indexable portion", excluding a portion of the web that has been called "the hidden web" [1] and is characterized as all the pages that normal users could eventually access, but automated agents such as the crawlers used by search engines do not. Certain pages are not indexable because they require special authorization. Others are **dynamic pages**, generated after the request has been made. Many dynamic pages are indexable, as the parameters for creating them can be found by following links. This is the case of, e.g. typical product catalogs in Web stores, in which there are links to navigate the catalog without the user having to pose a query.

The amount of information in the Web is certainly finite, but when a dynamic page leads to another dynamic page, *the number of pages can be potentially infinite*. Take, for example, a dynamic page which implements a calendar; you can always click on "next month" and from some point over there will be no more data items in the calendar; humans can be reasonably sure that it is very unlikely to find events scheduled 50 years in advance, but a crawler can not. There are many more examples of "crawler traps" that involve loops and/or near-duplicates (which can be detected afterwards, but we want to avoid downloading them).

In this work, we deal with the problem of capturing a relevant portion of the *dynamically generated content with known parameters*, while avoiding the download of too many pages. We are interested in knowing if a user will ever see a dynamically generated page. If the probability is too low, would a search engine like to retrieve that page? Clearly, from the Web site point of view the answer is yes, but perhaps from the search engine's point of view, the answer is no. In that case, our results are even more relevant. The answer in the case of the user's point of view is not clear *a priori*, as will depend on the result.

The main contributions of this paper are the models we propose for random surfing inside a Web site when the number of pages is **unbounded**. To do that, we take the tree induced by the Web graph of a site, and study it by levels. We analyze these models, focusing on the question of how "deep" users go inside a Web site and we validate these models using actual data from Web sites and link analysis with Pagerank. Our results help to decide when a crawler should stop, and to evaluate how much and how important are the non-crawled pages.

The next section outlines prior work on this topic, and the rest of this paper is organized as follows: in section 3, three models of random surfing in dynamic Web sites are presented and analyzed; in section 4, these models are compared with actual data from the access log of several Web sites. Section 5 concludes with some final remarks and recommendations for practical web crawler implementations.

2 Previous Work

Crawlers are an important component of Web search engines, and as such, their internals are kept as business secrets. Recent descriptions of Web crawlers include: Mercator [2], Salticus [3], WIRE [4], a parallel crawler [5] and the general crawler architecture described by Chakrabarti [6].

Models of random surfers as the one studied by Diligenti *et al.* [7] have been used for page ranking using the Pagerank algorithm [8], and for sampling the web [9]. Other studies about Web crawling have focused in crawling policies to capture high-quality pages [10] or to keep the search engine's copy of the Web up-to-date [11]. Link analysis on the Web is currently a very active research topic; for a concise summary of techniques, see a survey by Henzinger [12].

Log file analysis has a number of restrictions arising from the implementation of HTTP, specially caching and proxies, as noted by Haigh and Megarity [13]. *Caching* implies that re-visiting a page is not always recorded, and re-visiting pages is a common action, and can account for more than 50% of the activity of users, when measuring it directly in the browser [14]. *Proxies* implies that several users can be accessing a Web site from the same IP address. To process log file data, careful data preparation must be done [15], including the detection of sessions from automated agents [16].

The visits to a Web site have been modeled as a sequence of decisions by Huberman *et. al* [17, 18]; they obtain a model for the number of clicks that follows a Zipf's law. Levene *et al.* [19] proposed to use an absorbing state to represent the user leaving the Web site, and analyzed the lengths of user sessions when the probability of following a link increases with session length. Lukose and Huberman [20] also present an analysis of the Markov chain model of a user clicking through a Web site, and focus in designing an algorithm for automatic browsing, which is also the topic of a recent work by Liu *et al.* [21].

3 Random Surfer Models for a Web site with Infinite Number of Pages

We will consider a Web site as a set of pages under the same host name, and a *user session* as a finite sequence of page views in this Web site. The starting point of a user session does not need to be the page located at the root directory of the server, as some users may enter to the Web site following a link to an internal page.

The *page depth* of a page in a session is the shortest path from the start page through the pages seen during a session. This is not only a function of the Web site structure, this is the perceived depth during a particular session. The *session depth* is the maximum depth of a page in a session.

For random surfing, we can model each page as a state in a system, and each hyperlink as a possible transition; or we can use a simpler model in which we collapse multiple pages at the same level as a single node, as shown in Figure 1 (left and center). That is, the Web site graph is collapsed to a sequential list.

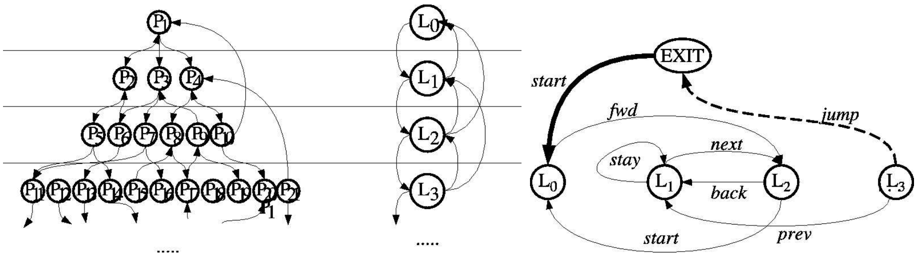

Fig. 1. Left: a Web site modeled as a tree. Center: the Web site modeled as a sequence of levels. Right: Representation of the different actions of the random surfer.

At each step of the walk, the surfer can perform one of the following actions, which we consider as atomic: go to the next level (action *next*), go back to the previous level (action *back*), stay in the same level (action *stay*), go to a different previous level (action *prev*), go to a different higher level (action fwd), go to the start page (action *start*) or jump outside the Web site (action *jump*). For action *jump* we add an extra node `EXIT` to signal the end of a user session (closing the browser, or going to a different Web site) as shown in Figure 1 (right). Regarding this Web site, after leaving, users have only one option: start again in a page with depth 0 (action *start*).

As this node `EXIT` has a single out-going link with probability 1, it does not affect the results for the other nodes if we remove the node `EXIT` and change this by transitions going to the start level L_0. Another way to understand it is that as this process has no memory, going back to the start page or starting a new session are equivalent, so actions *jump* and *start* are indistinguishable in terms of the resulting probability distribution for the other nodes. The set of atomic actions is $\mathcal{A} = \{next, start/jump, back, stay, prev, fwd\}$.

The probability of an action at level ℓ is $Pr(action|\ell)$. As they are probabilities $\sum_{action \in \mathcal{A}} Pr(action|\ell) = 1$. The probability distribution of being at a level at a given time is the vector $\mathbf{x}(t) = (x_0, x_1, \dots)$. When there exists a limit, we will call this $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}$.

In this article, we study three models with $Pr(next|\ell) = q \ \forall\ell$, i.e.: the probability of advancing to the next level is constant for all levels. Our purpose is to predict how far will a real user go into a dynamically generated Web site. If we know that, e.g.: $x_0 + x_1 + x_2 \geq 90\%$, then the crawler could decide to crawl just those three levels. The models we analyze were chosen to be as simple and intuitive as possible, though without sacrificing correctness. We seek more than just fitting the distribution of user clicks, we want to *understand and explain user behavior in terms of simple operations.*

3.1 Model A: back one level at a time

In this model, with probability q the user will advance deeper, and with probability $1-q$ the user will go back one level, as shown in Figure 2.

$Pr(next|\ell) = q$
$Pr(back|\ell) = 1 - q$ for $\ell \geq 1$
$Pr(stay|\ell) = 1 - q$ for $\ell = 0$
$Pr(start, jump|\ell) = 0$
$Pr(prev|\ell) = Pr(fwd|\ell) = 0$

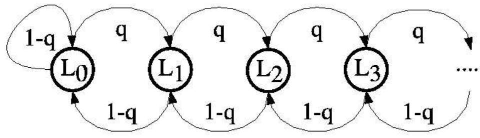

Fig. 2. Model A, the user can go forward or backward one level at a time.

A stable state $\mathbf{x}$ is characterized by $\sum_{i \geq 0} x_i = 1$ and:

$$x_i = q x_{i-1} + (1-q) x_{i+1} \qquad (\forall i \geq 1)$$
$$x_0 = (1-q) x_0 + (1-q) x_1$$

The solution to this recurrence is: $x_i = x_0 \left(\frac{q}{1-q}\right)^i \qquad (\forall i \geq 1)$.

If $q \geq 1/2$ then we have the solution $x_i = 0$, and $x_\infty = 1$ (that is, we have an absorbing state); which in our framework means that no depth can ensure that a certain proportion of pages have been visited by the users. When $q < 1/2$ and we impose the normalization constraint, we have a geometric distribution:

$$x_i = \left(\frac{1-2q}{1-q}\right) \left(\frac{q}{1-q}\right)^i$$

The cumulative probability of levels $0 \dots k$ is:

$$\sum_{i=0}^{k} x_i = 1 - \left(\frac{q}{1-q}\right)^{k+1}$$

3.2 Model B: back to the first level

In this model, the user will go back to the start page of the session with probability $1-q$. This is shown in Figure 3.

$Pr(next|\ell) = q$
$Pr(back|\ell) = 1 - q$ if $\ell = 1$, 0 otherwise.
$Pr(stay|\ell) = 1 - q$ for $\ell = 0$
$Pr(start, jump|\ell) = 1 - q$ for $\ell \geq 2$
$Pr(prev|\ell) = Pr(fwd|\ell) = 0.$

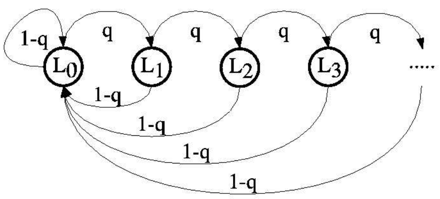

Fig. 3. Model B, the user can go forward one level at a time, or she/he can go back to the first level either by going to the start page, or by starting a new session.

A stable state $\mathbf{x}$ is characterized by $\sum_{i \geq 0} x_i = 1$ and:

$$x_0 = (1-q) \sum_{i \geq 0} x_i = (1-q)$$
$$x_i = q x_{i-1} \qquad (\forall i \geq 1)$$

As we have $q < 1$ we have another geometric distribution: $x_i = (1-q)q^i$. The cumulative probability of levels $0 \dots k$ is: $\sum_{i=0}^{k} x_i = 1 - q^{k+1}$.

Note that the cumulative distribution obtained with model A ("back one level") using parameter q_A, and model B ("back to home") using parameter q_B are equivalent if: $q_A = \frac{q_B}{1+q_B}$. So, as the distribution of session depths is equal, except for a transformation in the parameter q, we will consider only model B for charting and fitting the distributions.

3.3 Model C: back to any previous level

In this model, the user can either discover a new level with probability q, or go back to a previous visited level with probability $1-q$. If he decides to go back to a previously seen level, he will choose uniformly from he set of visited levels (including the current one), as shown in Figure 4.

A stable state $\mathbf{x}$ is characterized by $\sum_{i \geq 0} x_i = 1$ and:

$$x_0 = (1-q) \sum_{k \geq 0} \frac{x_k}{k+1}$$
$$x_i = q x_{i-1} + (1-q) \sum_{k \geq i} \frac{x_k}{k+1} \qquad (\forall i > 1)$$

$Pr(next|\ell) = q$
$Pr(back|\ell) = 1 - q/(\ell+1),\ \ell \geq 1$
$Pr(stay|\ell) = 1 - q/(\ell+1)$
$Pr(start, jump|\ell) = 1 - q/(\ell+1),$
$\ell \geq 2$
$Pr(prev|\ell) = 1 - q/(\ell+1),\ \ell \geq 3$
$Pr(fwd|\ell) = 0.$

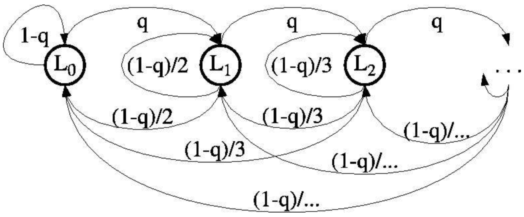

Fig. 4. Model C: the user can go forward one level at a time, and can go back to previous levels with uniform probability.

We can take a solution of the form: $x_i = x_0\,(i+1)\,q^i$. Imposing the normalization constraint, this yields: $x_i = (1-q)^2\,(i+1)\,q^i$. The cumulative probability of levels $0 \ldots k$ is: $\sum_{i=0}^{k} x_i = 1 - (2 + k - (k+1)\,q)\,q^{k+1}$.

3.4 Comparison of the Models

In terms of the cumulative probability of visiting the different levels, models A and B produce equivalent results except for a transformation of the parameters. Plotting the cumulative distributions for models B and C yields Figure 5. We can see that if $q \leq 0.4$, then in these models there is no need for the crawler to go past depth 3 or 4 to capture more than 90% of the pages a random surfer will actually visit, and if q is larger, say, 0.6, then the crawler must go to depth 6 or 7 to capture this amount of page views.

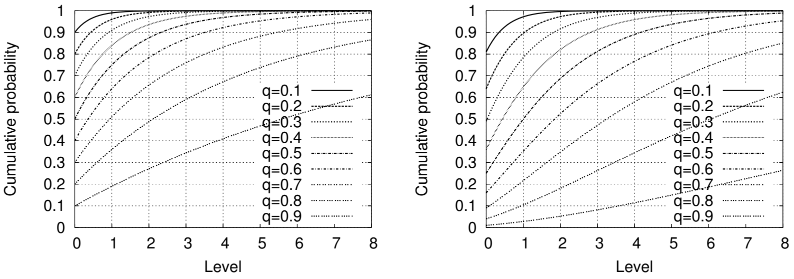

Fig. 5. Cumulative probabilities for models B (left) and C (right)

4 Data from user sessions in Web sites

We studied real user sessions on 13 different Web sites in the US, Spain, Italy and Chile, including commercial, educational, non-governmental organizations

Collection						Fit		
Code	Type	Country	Recorded sessions	Average page views	Root entry	Best model	q	Error
E1	Educational	Chile	5,500	2.26	84%	B	0.51	0.88%
E2	Educational	Spain	3,600	2.82	68%	B	0.51	2.29%
E3	Educational	US	71,300	3.10	42%	B	0.64	0.72%
C1	Commercial	Chile	12,500	2.85	38%	B	0.55	0.39%
C2	Commercial	Chile	9,600	2.09	32%	B	0.62	5.17%
R1	Reference	Chile	36,700	2.08	11%	B	0.54	2.96%
R2	Reference	Chile	14,000	2.72	22%	B	0.59	2.75%
O1	Organization	Italy	10,700	2.93	63%	C	0.35	2.27%
O2	Organization	US	4,500	2.50	1%	B	0.62	2.31%
OB1	Organization + Blog	Chile	10,000	3.73	31%	B	0.65	2.07%
OB2	Organization + Blog	Chile	2,000	5.58	84%	B	0.72	0.35%
B1	Blog	Chile	1,800	9.72	39%	C	0.79	0.88%
B2	Blog	Chile	3,800	10.39	21%	C	0.63	1.01%

Table 1. Characteristics of the studied Web sites and results of fitting the models. The number of user sessions does not reflect the relative traffic of the Web sites, as the data was obtained in different time periods. "Root entry" is the fraction of sessions starting in the home page.

and Web logs (sites in in which collaborative forums play a major role, also known as "Blogs"); characteristics of this sample, as well as the results of fitting models B and C to the data are summarized in Table 1.

We obtained access logs with anonymous IP addresses from these Web sites, and processed them to obtain user sessions, considering a session as a sequence of `GET` requests with the same `User-Agent` [22] and less than 30 minutes between requests [23]. We also processed the log files to discard hits to Web applications such as e-mail or content management systems, as they neither respond to the logic of page browsing, nor are usually accessible by Web crawlers. We expanded sessions with missing pages using the `Referrer` field of the requests, and considering all frames in a multi-frame page as a single page. Finally, we discarded sessions by Web robots [16] using known `User-Agent` fields and accesses to the `/robots.txt` file, and we discarded requests searching for buffer overflows or other software bugs.

As re-visits are not always recorded because of caching [14], data from log files *overestimates the depth at which users spent most of the time.* Figure 6 shows the cumulative distribution of visits per page depth to Web sites. At least 80%-95% of the visits occur at depth ≤ 4, and about 50% of the sessions include only the start page. The average session length is 2 to 3 pages, but in the case of Web logs, sessions tend to be longer. This is reasonable as Web postings are very short so Blog users view several of them during one session.

We fitted the models to the data from Web sites, as shown in Table 1 and Figure 8. In general, the curves produced by model B (and model A) are a better

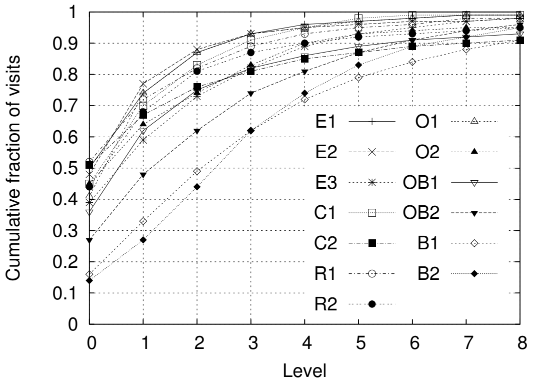

Fig. 6. Distribution of visits per level, from access logs of Web sites. E=educational, C=commercial, O=non-governmental organization, OB=Organization with on-line forum, B=Blog (Web log or on-line forum).

approximation to the user sessions than the distribution produced by model C, except for Blogs. The approximation is good for characterizing session depth, with error in general lower than 5%.

We also studied the empirical values for the distribution of the different actions at different levels in the Web site. We averaged this distribution across all the studied Web sites at different depths. The results are shown in Table 2, in which we consider all the Web sites except for Blogs.

Level	Observations	Next	Start	Jump	Back	Stay	Prev	Fwd
0	247985	**0.457**	–	**0.527**	–	0.008	–	0.000
1	120482	**0.459**	–	**0.332**	**0.185**	0.017	–	0.000
2	70911	**0.462**	**0.111**	**0.235**	**0.171**	0.014	–	0.001
3	42311	**0.497**	0.065	**0.186**	**0.159**	0.017	0.069	0.001
4	27129	**0.514**	0.057	**0.157**	**0.171**	0.009	0.088	0.002
5	17544	**0.549**	0.048	**0.138**	**0.143**	0.009	**0.108**	0.002
6	10296	**0.555**	0.037	**0.133**	**0.155**	0.009	**0.106**	0.002
7	6326	**0.596**	0.033	**0.135**	**0.113**	0.006	**0.113**	0.002
8	4200	**0.637**	0.024	**0.104**	**0.127**	0.006	0.096	0.002
9	2782	**0.663**	0.015	**0.108**	**0.113**	0.006	0.089	0.002
10	2089	**0.662**	0.037	0.084	**0.120**	0.005	0.086	0.003

Table 2. Average distribution of the different actions in user sessions, without considering Blogs. Transitions with values greater than 0.1 are shown in bold face.

We can see in Table 2 that the actions *next*, *jump* and *back* are the more important ones, which is in favor of models A (back one level) and model B (back to start level). We also note that $Pr(next|\ell)$ doesn't vary too much, and lies between 0.45 and 0.6. It increases as ℓ grows which is reasonable as a user that already have seen several pages is more likely to follow a link.

$Pr(jump|\ell)$ is higher than $Pr(back|\ell)$ for the first levels, and it is much higher than $Pr(start|\ell)$. About half of the user sessions involve only one page from the Web site. $Pr(start|\ell)$, $Pr(stay|\ell)$ and $Pr(fwd|\ell)$ are not very common actions.

5 Conclusions

The models and the empirical data presented lead us to the following characterization of user sessions: they can be modeled as a random surfer that either advances one level with probability q, or leaves the Web site with probability $1-q$. In general $q \approx 0.45-0.55$ for the first few levels, and then $q \approx 0.65-0.70$. This simplified model is good enough for representing the data for Web sites, but:

- We could also consider Model A (back one level at a time), which is equivalent in terms of cumulative probability per level, except for a change in the parameters. Based on the empirical data, we observe that users at first just leave the Web site while browsing (Model B), but after several clicks, they are more likely to go back one level (Model A).
- A more complex model could be derived from empirical data, particularly one that considers that q depends on ℓ. We considered that for our purposes, which are related to Web crawling, the simple model is good enough.
- Model C appears to be better for Blogs. A similar study to this one, focused only in the access logs of Blogs seems a reasonable thing to do since Blogs represent a growing portion of on-line pages.

In all cases, the models and the data show evidence of a distribution of visits which is strongly biased to the first few levels of the Web site. According to this distribution, more than 90% of the visits are closer than 4 to 5 clicks away from the entry page in most of the Web sites. In Blogs, we observed deeper user sessions, with 90% of the visits within 7 to 8 clicks away from the entry page.

In theory, as internal pages can be starting points, it could be concluded that Web crawlers must always download entire Web sites. In practice, this is not the case: if we consider the physical page depth in the directory hierarchy of a Web site, we observe that the distribution of surfing entry points per level rapidly decreases, so the overall number of pages to crawl is finite, as shown in Figure 7 (left).

Link analysis, specifically Pagerank, provides more evidence for our conclusions. We asked, what fraction of the total Pagerank score is captured by the pages on the first ℓ levels of the Web sites? To answer this, we crawled a large portion of the Chilean Web (.cl) obtaining around 3 million pages on April 2004, using 150 thousand seed pages that found 53 thousand Web sites. Figure 7 (right) shows the cumulative Pagerank score for this sample. Again, the first five levels

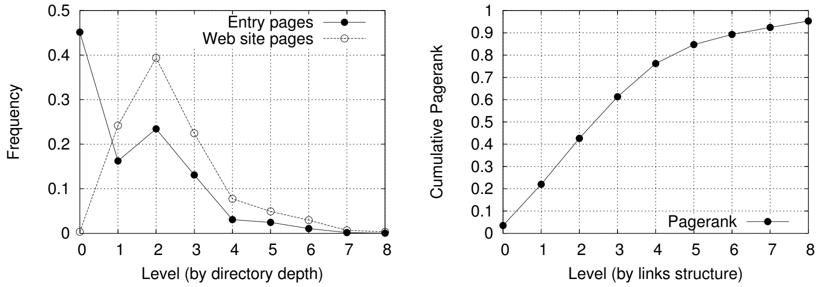

Fig. 7. Left: fraction of different Web pages seen at a given depth, and fraction of entry pages at the same depth in the studied Web sites, considering their directory structure. Right: cumulative Pagerank by page levels in a large sample of the Chilean Web.

capture 80% of the best pages. Note that the levels in this figure are obtained in terms of the global Web structure, considering internal and external links, not user sessions, as in the study by Najork and Wiener [10].

These models and observations could be used by a search engine, and we expect to do future work in this area. For instance, if the search engine's crawler performs a breadth-first crawling and can measure the ratio of new URLs from a Web site it is adding to its queue vs. seen URLs, then it should be able to infer how deep to crawl that specific Web site. The work we presented in this article provides a framework for that kind of adaptivity.

An interesting enhancement of the models shown here is to consider the contents of the pages to detect duplicates and near-duplicates. In our model, downloading a duplicate page should be equivalent to going back to the level at which we visited that page for the first time. A more detailed analysis could also consider the distribution of terms in Web pages and link text as the user browses through a Web site.

As the amount of on-line content that people, organizations and business are willing to publish grows, more Web sites will be built using Web pages that are dynamically generated, so those pages cannot be ignored by search engines. Our aim is to generate guidelines to crawl these new, practically infinite, Web sites.

References

1. Raghavan, S., Garcia-Molina, H.: Crawling the hidden web. In: Proceedings of the Twenty-seventh International Conference on Very Large Databases (VLDB), Rome, Italy, Morgan Kaufmann (2001) 129–138
2. Heydon, A., Najork, M.: Mercator: A scalable, extensible web crawler. World Wide Web Conference **2** (1999) 219–229
3. Burke, R.D.: Salticus: guided crawling for personal digital libraries. In: Proceedings of the first ACM/IEEE-CS joint conference on Digital Libraries, Roanoke, Virginia (2001) 88–89

4. Baeza-Yates, R., Castillo, C.: Balancing volume, quality and freshness in web crawling. In: Soft Computing Systems - Design, Management and Applications, Santiago, Chile, IOS Press Amsterdam (2002) 565–572
5. Cho, J., Garcia-Molina, H.: Parallel crawlers. In: Proceedings of the eleventh international conference on World Wide Web, Honolulu, Hawaii, USA, ACM Press (2002) 124–135
6. Chakrabarti, S.: Mining the Web. Morgan Kaufmann Publishers (2003)
7. Diligenti, M., Gori, M., Maggini, M.: A unified probabilistis framework for web page scoring systems. IEEE Transactions on Knowledge and Data Engineering **16** (2004) 4–16
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation algorithm: bringing order to the web. In: Proceedings of the seventh conference on World Wide Web, Brisbane, Australia (1998)
9. Henzinger, M., Heydon, A., Mitzenmacher, M., Najork, M.: On near–uniform url sampling. In: Proceedings of the Ninth Conference on World Wide Web, Amsterdam, Netherlands, Elsevier Science (2000) 295–308
10. Najork, M., Wiener, J.L.: Breadth-first crawling yields high-quality pages. In: Proceedings of the Tenth Conference on World Wide Web, Hong Kong, Elsevier Science (2001) 114–118
11. Cho, J., Garcia-Molina, H.: Synchronizing a database to improve freshness. In: Proceedings of ACM International Conference on Management of Data (SIGMOD), Dallas, Texas, USA (2000) 117–128
12. Henzinger, M.: Hyperlink analysis for the web. IEEE Internet Computing **5** (2001) 45–50
13. Haigh, S., Megarity, J.: Measuring web site usage: Log file analysis. Network Notes (1998)
14. Tauscher, L., Greenberg, S.: Revisitation patterns in world wide web navigation. In: Proceedings of the Conference on Human Factors in Computing Systems CHI'97. (1997)
15. Tanasa, D., Trousse, B.: Advanced data preprocessing for intersites Web usage mining. IEEE Intelligent Systems **19** (2004) 59–65
16. Tan, P.N., Kumar, V.: Discovery of web robots session based on their navigational patterns. Data Mining and Knowledge discovery **6** (2002) 9–35
17. Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M.: Strong regularities in world wide web surfing. Science **280** (1998) 95–97
18. Adar, E., Huberman, B.A.: The economics of web surfing. In: Poster Proceedings of the Ninth Conference on World Wide Web, Amsterdam, Netherlands (2000)
19. Levene, M., Borges, J., Loizou, G.: Zipf's law for web surfers. Knowledge and Information Systems **3** (2001) 120–129
20. Lukose, R.M., Huberman, B.A.: Surfing as a real option. In: Proceedings of the first international conference on Information and computation economies, ACM Press (1998) 45–51
21. Liu, J., Zhang, S., Yang, J.: Characterizing web usage regularities with information foraging agents. IEEE Transactions on Knowledge and Data Engineering **16** (2004) 566 – 584
22. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems **1** (1999) 5–32
23. Catledge, L., Pitkow, J.: Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems **6** (1995)

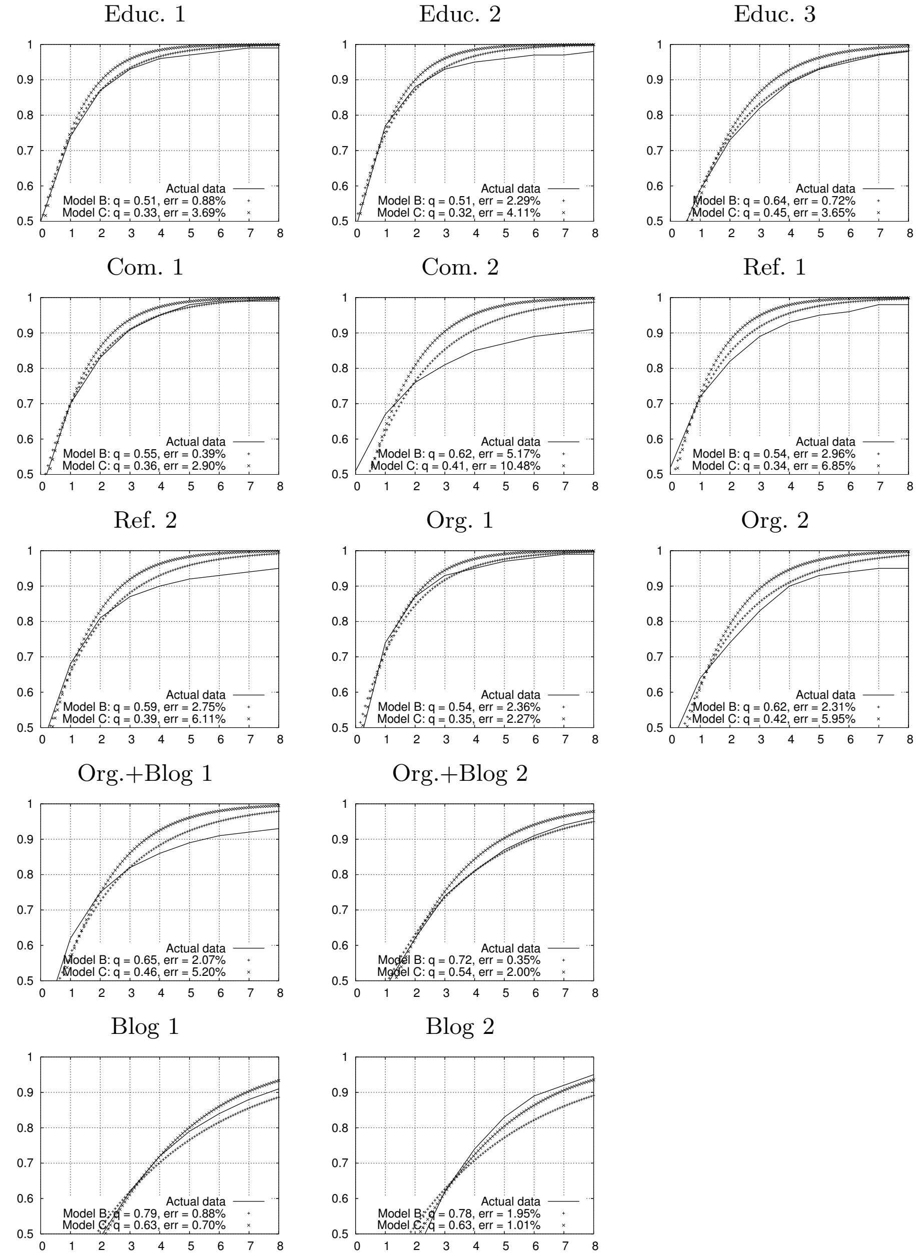

Fig. 8. Fit of the models to actual data, in terms of cumulative page views per level. Model B (back to start level), has smaller errors for most Web sites, except for Blogs. The asymptotic standard error for the fit of this model is 5% in the worst case, and consistently less than 3% for all the other cases. Note that we have zoomed into the upper portion of the graph, starting in 50% of cumulative page views.