Average-clicks: A New Measure of Distance on the World Wide Web

Yutaka matsuo (matsuo@miv.t.u-tokyo.ac.jp)
PRESTO, Japan Science and Technology Corporation and Graduate School of Engineering, University of Tokyo

Yukio Ohsawa (osawa@gssm.otsuka.tsukuba.ac.jp)
PRESTO, Japan Science and Technology Corporation and Graduate School of Business Science, University of Tsukuba

Mitsuru Ishizuka (ishizuka@miv.t.u-tokyo.ac.jp)
Graduate School of Information Science and Technology, University of Tokyo

Abstract. The pages and hyperlinks of the World Wide Web may be viewed as nodes and edges in a directed graph. In this paper, we propose a new definition of the distance between two pages, called *average-clicks*. It is based on the probability to click a link through random surfing. We compare the average-clicks measure to the classical measure of clicks between two pages, and show the average-clicks fits better to the users' intuition of distance.

Keywords: average-clicks, link structure, random surfer model, intuition of distance

1. Introduction

The World Wide Web provides considerable auxiliary information on top of the texts of the Web pages, such as its link structure. There have been a number of recent activities on how to utilize the link structure of the Web. Kleinberg distinguished between two types of Web sites which pertain to a certain search topic: *hubs* and *authorities*. The hub scores and authority scores are determined by an iterative procedure (Kleinberg et al., 1999). In order to find hubs and authorities, Lempel and Moran develop an algorithm called SALSA, which performs a random walk by alternately (a) going uniformly to one of the pages which links to the current page, and (b) going uniformly to one of the pages linked to by the current page (Lempel and Moran, 2000).

The Google[1] search engine uses the link structure for ranking Web pages, called PageRank (Brin and Page, 1998). A page has high rank if the sum of the ranks of its backlinks is high. And the rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to. PageRank is a global ranking of all Web pages, regardless of their contents, based solely on their location in the Web's graph structure.

Most of these works, which analyze the structure of the Web graph, assume the length of each link to be 1 (unit), and the clicks between two pages are counted to measure the distance. For example, Kumar et al. (1999) finds a bipartite core, which is a densely linked community consisting of a set of authorities and a set of hubs within 1 click. Albert et al. shows that two randomly chosen documents on the Web are on average 19 clicks away from each other (Albert et al., 1999). However, the distance measured by the number of clicks doesn't reflect well the users' intuition of distance. Some pages have incredibly large amount of links, while most pages have 10 or less links (Broder et al., 2000). For users, it requires a great effort to find and click a link among a large number of links than a link among a couple of links. If we count minimal clicks to measure the distance between two pages, the path is likely to include link collections, such as Yahoo![2] directories. To grab the users' intuition of distance correctly is essential for some research using link structure, e.g., identifying Web communities and focused crawling. A better approximation than 'clicks' is desirable to measure the distance of users' context.

In this paper, we propose a new definition of the distance between two pages, called *average-clicks* instead of the classical "clicks" measure. This measure reflects how many "average clicks" are needed from a page to another page. An average-click is one click among n links[3]. And two average-clicks is a distance of two successive clicks among n links for each, or one click among n^2 links. The average-click is defined on the probability for a "random surfer" to reach the page, based on the same idea as PageRank: A random surfer keeps clicking on successive links at random. The probability for a random surfer in page p to click one of the links in page p is considered as $1/OutDegree(p)$ in this model, (ignoring the damping factor). We annotate the link in page p with the length of $-log_n(1/OutDegree(p))$, so that summing lengths is akin to multiplying probabilities. An average-click is a unit distance of this measure.

If we measure the distance by average-clicks, the path through a large link collection can be considered long even if it takes only a couple of clicks. On the contrary, the path in a line of pages is considered short even if many clicks are necessary. This fits very well to the users' intuition of distance. We show by questionnaires that our average-clicks is a better model to approximate the users' intuition than the classical clicks measure.

In the following section, the definition of average-clicks is explained in detail. In Section 3 and 4, we show some examples and a questionnaire data analysis on the user's concept of distance. We discuss related

works and the possible application of average-clicks in Section 5, and conclude the paper.

2. Average-clicks

When analyzing the Web as a graph, we are confronted by the diversity of the links. There are not only topic related links, but also intra domain links, commercial/sponsor links, and so on. Some pages have more than a hundred of links, while others have a few or no links. The variety is so wide that we want to weight these links by some means. Here we define the length of a link using only the number of the links in a page, inspired by the PageRank algorithm.

PageRank makes a probability distribution over Web pages, based on the simple idea that a "random surfer" keeps clicking on successive links at random. The probability to click each link in page p is $\alpha/OutDegree(p)$, where α is a damping factor and $OutDegree(p)$ is the number of links page p has. In probability $1-\alpha$, a random surfer jumps to a random Web page. Although a user on the Web does not actually act as a 'random surfer,' this model gives us very effective approximation which can be seen as probabilistic distribution of users on the Web pages (Page et al., 1998). Following Page et al. (1998) α is usually set to be 0.85, however, we set $\alpha = 1$ below for simplicity[4].

We annotate a link with the length as negative logarithm of probability, so that summing lengths is akin to multiplying probabilities.

Definition 1
A length of a link in page p is defined as

$$-log_n(\alpha/OutDegree(p)).$$

We set the base of the logarithm n to be 7 in this paper, due to the fact that the average page has roughly seven hyperlinks to other pages[5] (Bharat and Broder, 1998).

Definition 2
An *average-click* is a unit of a length defined by Definition 1 where $\alpha = 1.0$ and $n = 7$.

We can set $\alpha < 1.0$ if we want to define the length of a hyperlink in the page with one hyperlink to be more than 0. One candidate is to set $\alpha = 0.85$ following (Page et al., 1998), but we can't find a significant difference between $\alpha = 0.85$ and $\alpha = 1$ in our experiment. We can also set n to be other value, e.g., 2, e or 10. For example, if we set $n = 2$, then the value shows amount of information (measured by "bits").

In this paper, n is set to be 7 in order the value to represent "how many clicks in an average page is needed from one page to another." We show in the following section that the average-clicks has stronger correlation with the users' percieved distance than clicks measure. Because correlation is irrelevant to magnitude of the coefficient (i.e., the base of the logarithm here), our claim is not affected even if we don't use $n = 7$; if we use another value than 7, only semantics will changes.

The distance between two pages p and q is defined by the shortest path. From a probabilistic point of view, this is equivalent to focus only on the path with the largest probability for a random surfer to get from page p to page q.

Definition 3
The *distance* from p to q is the sum of the length of the shortest path from p to q.

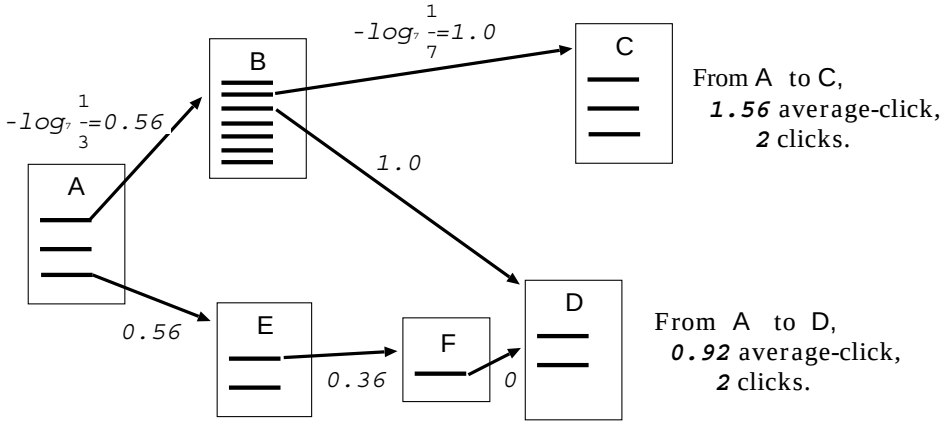

Figure 1. Average-clicks and clicks.

Fig. 1 illustrates some pages and the links between them. Page A has three links, thus the length of each link is $-\log_7(1/3) \approx 0.56$ average-click. As page B has seven links, the length is each 1 average-click. Summing 0.56 and 1, the distance from A to C is 1.56 average-click. In the case of page D, there is two paths from page A to D. The average-clicks is smaller in the lower path, though it takes three clicks. The shortest path in terms of average-clicks is the lower path, while the path with minimal clicks is the upper path. Note that if a page has only one link, as page F, the length of the link is 0 average-click[6].

This model offers a very good approximation to our intuitive concept of distance between Web pages. For example, Yahoo! top-page has currently more than 180 links. In our definition, the length from the top-page to each sub-page is very far, as the upper path in Fig. 1. On

the other hand, the path length by the local relation, such as the link to one's friends or the link to one's interests, is estimated rather short, as in the lower path of the figure. Intuitively we think the path through the Yahoo! top-page is longer than the path along the acquaintance chain with the same clicks. In our model, page C is more distant from page A than page D, and this fits very well to our intuition.

```
function FIND_SHORTEST_PATH (starting_url, target_url, d_thre)
  α ← 1.0,    n ← 7 ;
  url ← starting_url ;
  d(starting_url) ← 0 ;
  url_queue ← empty ;
  while (url ≠ target_url)
    page ← CRAWL_PAGE(url) ;
    url_list ← EXTRACT_URLS(page) ;
    n_cur ← # of current elements of url_list ;
    if n_cur > 0 then
      d_cur ← d(url) − log_n(α/n_cur) ;
      if d_cur < d_thre then
        for each u in url_list
          if d(u) is not defined then
            d(u) ← d_cur ;
            ENQUE(url_queue, u) ;
          else
            if d(u) > d_cur then
              d(u) ← d_cur ;
        sort url_queue by d ;
    if url_queue is empty then return failure
    url ← DEQUEUE(url_queue) ;
  endwhile
  return d(target_url)
```

d_{thre}	:	the range of the search space.
CRAWL_PAGE(url)	:	crawl url and return the contents.
EXTRACT_URLS($page$)	:	extract links from $page$.
ENQUE ($queue, element$)	:	append element at the end of queue.
DEQUEUE ($queue$)	:	remove the element at the beginning of queue and return it.

Figure 2. The best first search for the shortest path.

3. Case Study

In this section, we show some examples of the distance between two pages by the average-clicks measure. We first implement the best-first algorithm to search the shortest path from a starting url to a target url, as shown in Fig.2. This best-first algorithm is the similar algorithm to the breadth-first algorithm for the clicks measure. Najork et al. study what order a crawler should visit the URLs to obtain more important pages first, and showed that performing the crawl in breadth-first order works well (Najork and Wiener, 2001).

Table I shows an example of the distance from one of the author's homepage[7]. This homepage, stated below as page a, is located on the server at Tokyo University in Japan. The results showed the following:

- The search is not trapped into the link collection.
- The distance by average-clicks seems to fit well to our intuitive concept of distance. In other words, pages familiar to the author of page a are estimated to be near, and unfamiliar pages are estimated to be distant.
- The shortest path is very informative for the author in that it provides the indirect relation of two pages.

For example, the distance to one of the author's colleagues or Yahoo! is small, and they are very familiar to the author. The IJCAI homepage is more distant than the JSAI homepage. In fact, we participate in JSAI events more. The distance to WI-2001 is very far now, however, it might get shorter in the future for the very reason that we participated WI-2001.

4. Evaluation

4.1. Evaluation by users' feeling of distance

How can we evaluate that the average-clicks fits well to user's intuition of distance? We don't try to evaluate by analyzing textual contents of Web pages because we want to reveal the intuitive distance, rather than the similarity of contents. For example, a page for one's research topic and a page for the related conference have considerably different sets of terms, however they can be in small intuitive distance. The perceived distance is subjective, thus we try to evaluate by questionnaires. This section is devoted to show a preliminary report on the quantitative evaluation using questionnaires.

Table I. The distance measured by average-clicks from page *a*.

To URL Shortest path	Cumulative distance (average-clicks)
One of the author's colleagues	
http://www.miv.t.u-tokyo.ac.jp/~matumura/	1.62
http://www.miv.t.u-tokyo.ac.jp/JAICO/	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
Yahoo! (Japanese site)	
http://www.yahoo.co.jp/	3.02
http://www.geocities.co.jp/Athlete-Athene/6353/whatsnew.html	2.67
http://www.geocities.co.jp/Athlete-Athene/6353/	1.13
http://www.miv.t.u-tokyo.ac.jp/ matsuo	0.0
Japanese Society of Artificial Intelligence homepage	
http://www.nacsis.ac.jp/jsai/	4.69
http://www.miv.t.u-tokyo.ac.jp/~yabuki/	2.54
http://www.miv.t.u-tokyo.ac.jp/member/present-mem.htm	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
International Joint Conference on AI homepage	
http://ijcai.org/	5.39
http://w3.sys.es.osaka-u.ac.jp/~osawa/AIlinks.html	3.33
http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa	1.97
http://www.miv.t.u-tokyo.ac.jp/~matumura/research.html	1.62
http://www.miv.t.u-tokyo.ac.jp/JAICO/	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
WI-2001 homepage	
http://kis.maebashi-it.ac.jp/wi01	10.40
http://internet.aist-nara.ac.jp/research/security/	8.14
http://iplab.aist-nara.ac.jp/research.html.en	7.06
http://iplab.aist-nara.ac.jp/	5.80
http://shika.aist-nara.ac.jp/	4.13
http://www.miv.t.u-tokyo.ac.jp/~santi/oohm.html	2.54
http://www.miv.t.u-tokyo.ac.jp/member/present-mem.htm	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0

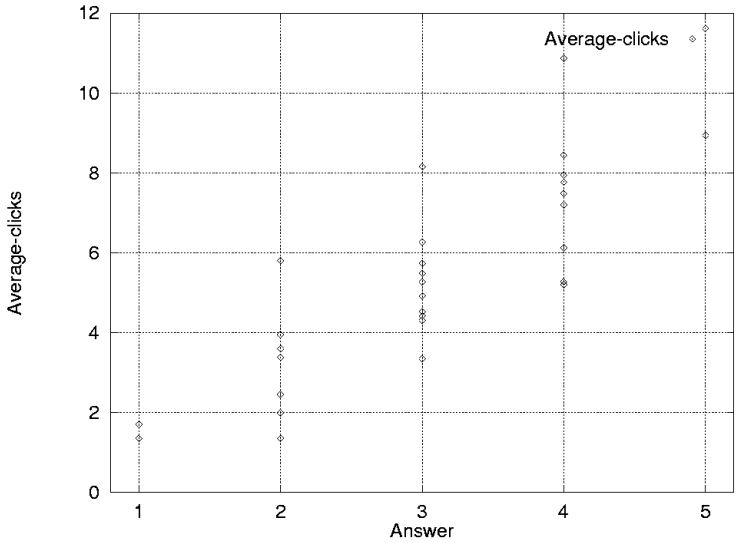

Figure 3. Scatter plot of answers and average-clicks by participant 1.

We asked 13 participants to rank the pages according to their perceived familiarity. First we pick up 30 pages randomly which we can obtain within a few clicks from each participant's homepage. Then, we asked him/her to answer how familiar each URL of the page is, without providing the contents of the pages or any distance measures. Answers to the questions were made on a 5-point Likert-scale from 1 (very familiar) to 5 (very distant). After the questionnaires, we compare users' ratings with the distance by clicks/average-clicks.

Fig. 3 and 4 shows the scatter plot of the results by participant 1. We can see very clearly that the rating is correlated with the average-clicks measure. On the other hand, the classical clicks measure doesn't seem to have a strong correlation with the ratings. The correlation coefficients of 13 participants are shown in Table II; if the correlation coefficient is close to 1, there is a strong positive correlation between two sets of data, and if the correlation coefficient is 0, there is no relationship. We also show the significant level at which a null hypothesis is rejected that the correlation coefficients of clicks and average-clicks are the same.

Average-clicks have generally stronger correlation with the users' ratings. In none of 13 cases, the correlation of clicks exceeds the correlation of average-clicks. Although in more than half the cases, we can't reject the hypothesis at a significance level of 5%. (Note that to validate the difference of two correlation coefficients, say 0.400 and 0.500 in a significance level of 5 %, we have to ask the participant to evaluate more than 150 URLs, which is overly burdensome.) However we can at

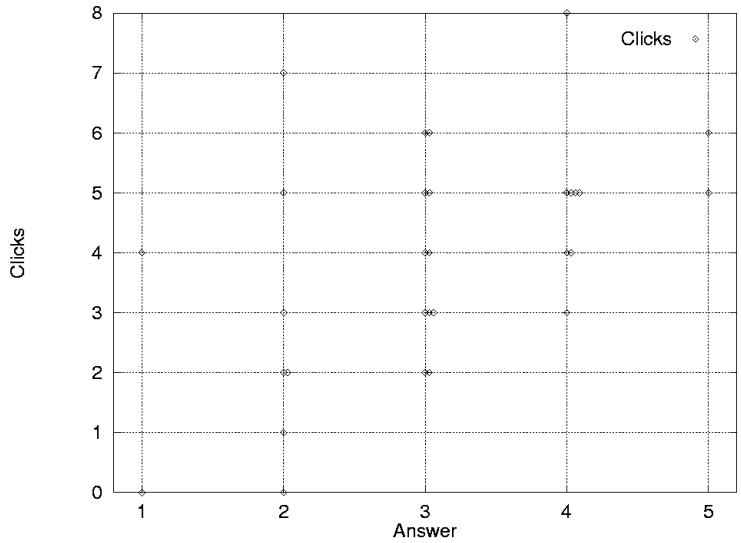

Figure 4. Scatter plot of answers and clicks by participant 1.

least conclude that it is likely that the average-clicks fits better to the users' rating than the clicks measure.

4.2. Evaluation by contents impression

We made another questionnaire which evaluates the users' intuition of distance by means of a different set of operationalizations of distance. Showing the participant the whole contents of the page for each 30 URL, we asked the following three questions:

- *understandability*; "How easy is this page to understand?" from 1 (very difficult) to 5 (very easy),
- *informativeness*; "How much does this page include what you don't know?" from 1 (very little) to 5 (very much),
- *interestingness*; "How interesting is this page?" from 1 (very dull) to 5 (very interesting).

Table III shows the average correlation coefficient of participants. Understandability has a negative correlation to the distance, that is, if the page is further, it gets more difficult to understand the contents. Informativeness has a positive correlation to the distance, that is, if the page is further, it is likely to be more informative. However, interestingness has almost nothing to do with distance. For one reason, each participant has different attitude towards novelty; some people find new

Table II. The correlation coefficient of participants' rating and clicks/average-clicks.

Participant	Correlation coefficient		Significance level
	Clicks	Average-clicks	of difference
1	0.524	0.836	1%
2	0.325	0.804	1%
3	0.471	0.685	5%
4	0.268	0.572	5%
5	0.641	0.674	above 5%
6	0.696	0.715	above 5%
7	0.517	0.699	above 5%
8	0.490	0.569	above 5%
9	0.499	0.528	above 5%
10	0.435	0.512	above 5%
11	0.358	0.436	above 5%
12	0.116	0.400	above 5%
13	0.302	0.367	above 5%
Average	0.434	0.600	—

Table III. Average correlation coefficient to understandability, informativeness and interestingness.

	clicks	average-clicks
understandability	-0.320	-0.404
informativeness	0.279	0.462
interestingness	-0.174	-0.174

things interesting, while others find familiar topics more interesting. Participants judge interestingness by considering many factors, such as understandability, informativeness, usefulness, quality, and so on.

Understandability and informativeness have stronger correlation with the average-clicks than the clicks measure. This is further supporting evidence of the average-clicks being a better measure of intuitive distance than the clicks measure. The average-clicks approximates the distance between the context of the participant homepage and the other page.

5. Discussion

Average-clicks is not always better than the clicks measure. In some cases, clicks performs better. For example, consider two professors in a university have their homepages, and both have a hyperlink to the university. Although the intuitive distance from the professors to the university should be the same (in that both are the professors of the university), the distance measured by average-clicks can be different. However, in most cases the average-clicks measure fits better than the clicks measure; if a homepage has a hyperlink to the university with smaller number of hyperlinks, then the intuitive distance (i.e., how familiar s/he feels about the university) is likely to be shorter. Of course this is not true for all the cases, but in total we showed that average-clicks has stronger correlation with perceived distance than clicks measure.

The evaluation in Section 4 is not inclusive; we can only support our conclusion from homepage owners' point of view. Besides individual homepages, there are Web pages made by companys, universities, Gorverment, and so on. Moreover, there are many types of Web pages including portal sites, news sites, link collections, communities, and so on. Further research is needed to show the usefulness of average-clicks measure for various kinds of Web pages.

The Web is an example of a social network. Social network theory is concerned with properties related to connectivity and distance in graphs (Chakrabarti, 2000). In Mendelzon and Rafiei (2000), the "reputation" of a Web page is mined from the link structure and textual information; for example, we can find `www.gamelan.com` has a reputation for "Java." Adamic and Adar (pear) shows a users' homepage and mailing lists can be used to predict relationships between individuals. Our research also reveals a part of a social network, by a simple measure.

In Chakrabarti et al. (1998), the weight of a link is defined by referring to the text of the page; if the text in the vicinity of the "href" contains text descriptive of the topic at hand, the weight of the link is increased. Bharat and Henzinger have invented another way to integrate textual contents with the link structure (Bharat and Henzinger, 1998). They model each page according to the "vector space" model, and prune the pages whose corresponding term vectors are "outliers" compared with other pages at one click away from the query result page. These weighing/pruning algorithms require the text analysis of a page, while our average-clicks measure requires only the number of links.

The average-clicks measure is another usage of the probability distribution by a random surfer model. To transform the probability into the length of a link, we can imagine more precisely the structure of the graph. This type of length (or cost) assignment is very common in the context of cost-based abduction, where finding the MAP (maximum a posteriori probability) solution is equivalent to finding the minimal cost explanation for a set of facts (Charniak and Shimony, 1994).

Many researchers now employ clicks as the measure of distance, however it seems reasonable to use average-clicks instead. For example, when finding a community on the Web, the general topics pages are likely to be included (Bharat and Henzinger, 1998). However, employing the average-clicks measure, the general topics pages are considered to be distant and can be filtered out, because such pages have usually many links. (According to Bharat and Henzinger (1998), the most-highly ranked authorities and hubs tend to be about general topics rather than the original topic.) Fetching the neighboring pages is a common procedure in many algorithms (Kosala and Blockeel, 2000). We should fetch the pages within a given threshold of average-clicks, not within a given threshold of clicks. A given threshold of clicks means sometimes an incredibly large range of the search. The average-clicks measure provides a good justification of the practical search, such as "if there are few links, fetch the pages, but if there are many links, give up." In the focused crawling, search in the breadth-first order as well as the PageRank order is shown to be a good strategy (Najork and Wiener, 2001; Cho et al., 1998). Our finding that the distant page means the distant context also supports the effectiveness of these strategies.

The classical clicks measure is intuitively understandable for all Internet users, while the distance based on the probability is relatively difficult to understand. That's why we bring semantics by setting the base of the logarithm to the average number of links in a page: The distance shows how many "average clicks" are needed from one page to another page.

6. Conclusion

In this paper, we have proposed a new measure, called average-clicks, and shown by questionnaires that our average-clicks is a better model to approximate the users' intuition than the classical clicks measure. The experiment is of course not conclusive, and it represents only a first step in the evaluation of effectiveness of the average-clicks. For example, we can use Web logs to evaluate the usefulness of average-clicks instead of questionnaires. However we are convinced that at least in some cases

we can use average-clicks rather than clicks measure. We will further investigate on how well our model fits for different kinds of Web pages. By modeling the Web structure more precisely, many research fields will benefit from finding communities to customized browsers.

Notes

[1] `http://google.com`

[2] `http://www.yahoo.com`

[3] In this paper, we set n to be 7 due to the fact that the average page has roughly seven hyperlinks to other pages.

[4] We have tried another value including 0.85 in the evaluation, but we don't find big differences.

[5] More recent survey shows the average page has 1 external link and 4 internal links (Murray and Moore, 2000).

[6] If we set α to be less than 1.0, the length of the link is more than 0.

[7] `http://www.miv.t.u-tokyo.ac.jp/~matsuo`

References

Adamic, L. A. and E. Adar: to appear, 'Friends and Neighbors on the Web'. URL://www.hpl.hp.com/shl/papers/web10/fnn.pdf.

Albert, R., H. Jeong, and A. Barabási: 1999, 'Diameter of the World-Wide Web'. *Nature* **401**(6749).

Bharat, K. and A. Broder: 1998, 'A technique for measuring the relative size and overlap of public Web search engines'. In: *Proc. 7th WWW Conf.*

Bharat, K. and M. R. Henzinger: 1998, 'Improved algorithms for topic distillation in a hyperlinked environment'. In: *Proc. 21st ACM SIGIR Conf.* pp. 104–111.

Brin, S. and L. Page: 1998, 'The Anatomy of a Large-Scale Hypertextual Web Search Engine'. In: *Proc. 7th WWW Conf.*

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener: 2000, 'Graph structure in the web'. In: *Proc. 9th WWW Conf.*

Chakrabarti, S.: 2000, 'Data mining for hypertext: A tutorial survey'. *SIGKDD Explorations* **1**(2), 1–11.

Chakrabarti, S., B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg: 1998, 'Automatic resource compilation by analyzing hyperlink structure and associated text'. In: *Proc. 7th WWW Conf.*

Charniak, E. and S. E. Shimony: 1994, 'Cost-based abduction and MAP explanation'. *Artificial Intelligence* **66**, 345–374.

Cho, J., H. Garcia-Molina, and L. Page: 1998, 'Efficient Crawling Through URL Ordering'. In: *Proc. 7th WWW Conf.*

Kleinberg, J. M., R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins: 1999, 'The Web as a graph: measurements, models, and methods'. In: *Proc. International Conf. on Combinatorics and Computing.*

Kosala, R. and H. Blockeel: 2000, 'Web mining research: A survey'. *ACM SIGKDD Explorations* **1**(2), 1–15.

Kumar, S. R., P. Raghavan, S. Rajagopalan, and A. Tokins: 1999, 'Trawling the web for emerging cyber communities'. In: *Proc. 8th WWW Conf.*

Lempel, R. and S. Moran: 2000, 'The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect'. In: *Proc. 9th WWW Conf.*

Mendelzon, A. O. and D. Rafiei: 2000, 'What do the Neighbours Think? Computing Web Reputations'. *IEEE Data Engineering Bulletin* **23**(3), 9–16.

Murray, B. and A. Moore: 2000, 'Sizing the Internet'. White paper, Cyveillance, Inc. (http://www.cyveillance.com).

Najork, M. and J. L. Wiener: 2001, 'Breadth-first search crawling yields high-quality pages'. In: *Proc. 10th WWW Conf.*

Page, L., S. Brin, R. Motwani, and T. Winograd: 1998, 'The PageRank Citation Ranking: Bringing order to the Web'. Technical report, Stanford University.

Address for Offprints:
Ishizuka lab., Information and Communication Engineering, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan