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

Yutaka Matsuo[1][2], Yukio Ohsawa[2][3], and Mitsuru Ishizuka[1]

[1] University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, JAPAN,
matsuo@miv.t.u-tokyo.ac.jp,
WWW home page: http://www.miv.t.u-tokyo.ac.jp/~matsuo/
[2] TOREST, Japan Science and Technology Corporation,
Tsutsujigaoka 2-2-11, Miyagino-ku, Sendai, Miyagi, 983-0852 Japan,
[3] University of Tsukuba, Otsuka 3-29-1, Bunkyo-ku, Tokyo 113-0012, JAPAN

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 average-clicks fits better to the users' intuitions of distance.

1 Introduction

The World Wide Web provides considerable auxiliary information on top of the text of the Web pages, such as its link structure. There has been a fair amount of recent activity on how to exploit the link structure of the Web. Kleinberg distinguished between two types of Web sites which pertain to a certain search topic: *hubs* and *authorities*. A good hub is a page that points to many good authorities and a good authority is a page that is pointed to by many good hubs [8]. The hub scores and authority scores are determined by an iterative procedure. The pages with the highest scores are returned as hubs and authorities for the search topic.

The Google[1] search engine uses the link structure for ranking Web pages, called PageRank [4]. 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 content, 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, [8] finds the bipartite core, which is a densely linked community consisting of a set of authorities and a set of hubs within 1 click. [1] shows that two randomly chosen documents on the web are

[1] http://google.com

on average 19 clicks away from each other. 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 [5]. 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 a minimal clicks to measure the distance between two pages, the path is likely to include link collections, such as Yahoo![2] directories.

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 lines 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, 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 4, 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 classify 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

[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.

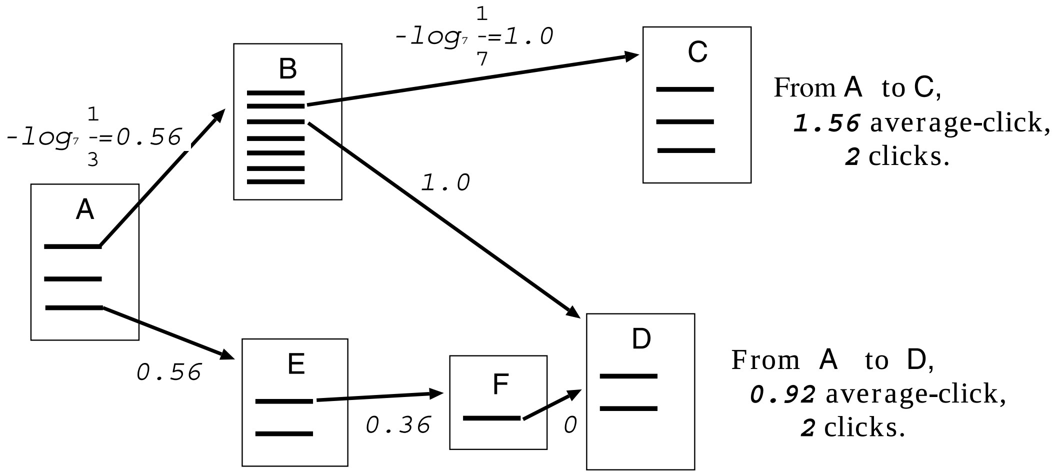

Fig. 1. Average-clicks and clicks.

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. Following [10] 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 [2][5]. We call a unit of the length an *average-click*.

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

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

[4] We are aware that setting $\alpha = 1$ means that the users always click on a page. A more realistic assumption is that there is some probalility α of following a link. However, since we don't have enough statistical results yet to decide α, we set simply $\alpha = 1$.

[5] The latest survey shows the average page has 1 external link and 4 internal links [9].

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.

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.

3 Case Study and Experimental Results

3.1 Examples of the distance

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 page s (stated as *start page*) to page t (stated as *target page*), as shown in Fig.2.

```
function SEARCH_SHORTEST_PATH ($start\_page$, $target\_page$, $d_{thre}$)
  $\alpha \leftarrow 1.0, \quad n \leftarrow 7.$
  $list \leftarrow$ ADD_LIST($start\_page, empty$), $d(start\_page) \leftarrow 0.$
  $p \leftarrow start\_page$
  while $p \neq target\_page$
    Fetch page $p$ and extract links which points to page $p_k$ $(k = 1, \ldots, n_p)$
    for $k \leftarrow 1$ to $n_p$
      $d(p_k) \leftarrow d(p) - log_n(\alpha / n_p)$
      if $d(p_k) > d_{thre}$ then next
      $list \leftarrow$ ADD_LIST($p_k, list$)
    end
    if $list$ is empty return failure
    $p \leftarrow$ CHOOSE_MINIMAL($list, d$)
  end
  return $d(target\_page)$
```

ADD_LIST($a, list$) is a function which add a to $list$.
CHOOSE_MINIMAL($list, d$) is a function which choose $a \in list$ minimizing $d(a)$.
d_{thre} is the range of the search space.

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

Table 1. The distance measured by average-clicks from page *a*.

To	
URL	Cumulative distance
Shortest path	(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

Table 1 shows an example of the distance from one of the author's homepage. This homepage, "`www.miv.t.u-tokyo.ac.jp/~matsuo`," 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 are submitting this paper to WI-2001.

3.2 Evaluation by questionnaires

This section shows a preliminary report on the quantitative evaluation using questionnaires. We asked five 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 compared the rating with the distance measure of clicks and 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 statistical results of five participants are shown in Table 2, which shows correlation coefficients: 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 can see from the table that the average-clicks have stronger correlation with the users' rating.

Table 2. The correlation coefficient of participants' rating and clicks/average-clicks.

Participant	Clicks	Average-clicks
1	0.524	0.836
2	0.696	0.715
3	0.517	0.699
4	0.325	0.804
5	0.471	0.685

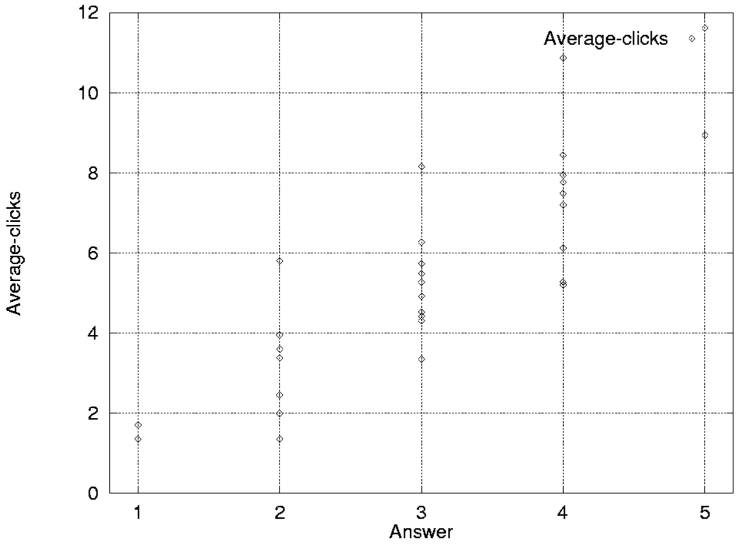

Fig. 3. Scatter plot of answers and average-clicks by participant 1.

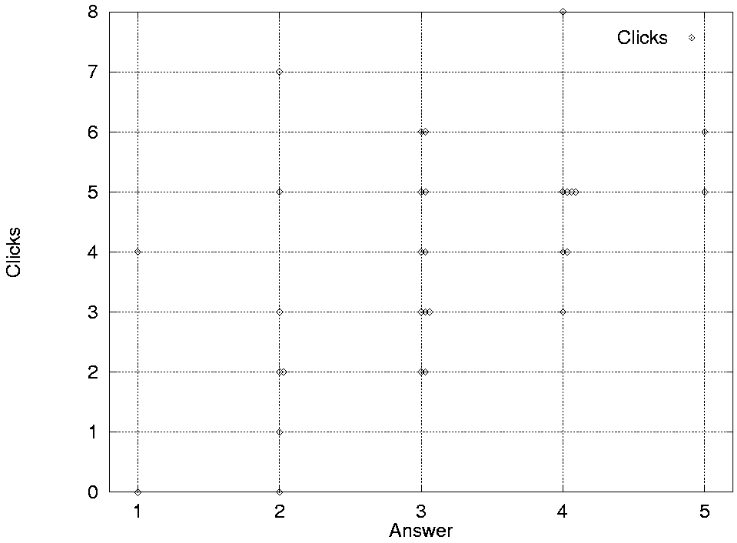

Fig. 4. Scatter plot of ansers and clicks by participant 1.

4 Discussion

In [6], 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. This weighing algorithm requires 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 [7].

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 [3]. 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[6]. Fetching the neighboring pages is a common procedure in many algorithms. 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. 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."

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.

5 Conclusion

In this paper, we have proposed a new measure, called average-clicks, and evaluate it by measuring the users' intuition of distance. By modelling the Web structure more precisely, many research fields will benefit from search engines to customized browsers. One of our future works is to estimate the value of a link using the average-clicks measure.

References

1. L. A. Adamic. The small world web. In *Proc. ECDL'99*, pages 443–452, 1999.

[6] The power law distribution on the Web says that the number of nodes with high out degrees is much less than the number of nodes with low degrees[5]. However we can't ignore these pages with high out degrees because they sometimes recieves many links.

2. K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proc. 7th WWW Conf.*, 1998.
3. K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. 21st ACM SIGIR conf.*, pages 104–111, 1998.
4. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th WWW Conf.*, 1998.
5. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proc. 9th WWW Conf.*, 2000.
6. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th WWW Conf.*, 1998.
7. E. Charniak and S. E. Shimony. Cost-based abduction and MAP explanation. *Artificial Intelligence*, 66:345–374, 1994.
8. J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: measurements, models, and methods. In *Proc. of the International Conference on Combinatorics and Computing*, 1999.
9. B. Murray and A. Moore. Sizing the internet (a white paper). White paper, Cyveillance, Inc., 2000. (http://www.cyveillance.com).
10. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. In *Techinical Report*, 1998. (http://www-db.stanford.edu/backrub/pageranksub.ps).