Discovering Hidden Relation behind a Link

Yutaka Matsuo[1,3] Yukio Ohsawa[2,3] Mitsuru Ishizuka[1]
[1] *Graduate School of Engineering, University of Tokyo,*
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
[2] *Graduate School of Systems Management, University of Tsukuba,*
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan
[3] *TOREST, Japan Science and Technology Corporation*
2-2-11 Zakurooka, Miyagino-ku, Sendai, Miyagi, 983-0852 Japan

Abstract. If a page has a link which points to another page, we can see a direct relation between two pages. However, what if the link isn't there? Is there still a path between two pages? In this paper, we propose an algorithm to show the second shortest path and reveal the hidden relation between two pages. The length of the second shortest path indicates how far the link bridges.

1 Introduction

Recently, there have been a number of algorithms proposed for analyzing hypertext link structures on the World Wide Web. Among them, Kleinberg's Hyperlink Induced Topic Search (HITS) [4] and the PageRank algorithm [3] underlying Google[1] are the most popular and refined algorithms.

These algorithms determine the best "authorities" for a given topic or query by ranking pages. However, few works have been done for ranking links. In this paper, the alternativeness of a link is discussed, based solely on the structure of the Web graph. For example, Yahoo![2] is one of the most useful pages, however the link to Yahoo! is so common that we can reach Yahoo! through any path on the Web. In this sense, the alternativeness of a link to Yahoo! can be estimated relatively high. In the year 2000, when the Google Japanese cite started a searvice, a link to Google was rare and considered to be valuable. Nowaday we see many links to Google, thus the value of one link to Google is getting low.

In our definition, a link is considered valuable if it prevents us from having to visit many pages to eventually reach a certain page, or if it enables us to reach a certain page in the first place. In short, a valuable link is a *shortcut*. The importance of a shortcut has been argued in the context of *small world* topology [5], where nodes are highly clustered yet the path length between them is small. Our justification is that as the Web is a small world [1], the shortcut must play an important role on the Web.

The range of a link is measured by the length of the second shortest path between two pages connected by the link. Seeking the second shortest path provides not only the alternativeness of the link, but also the hidden relation between the two pages connected by the link. Sometimes, the two pages have a surprising relation in the absence of the link. It may be a discovery of a hidden relation behind a link.

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

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

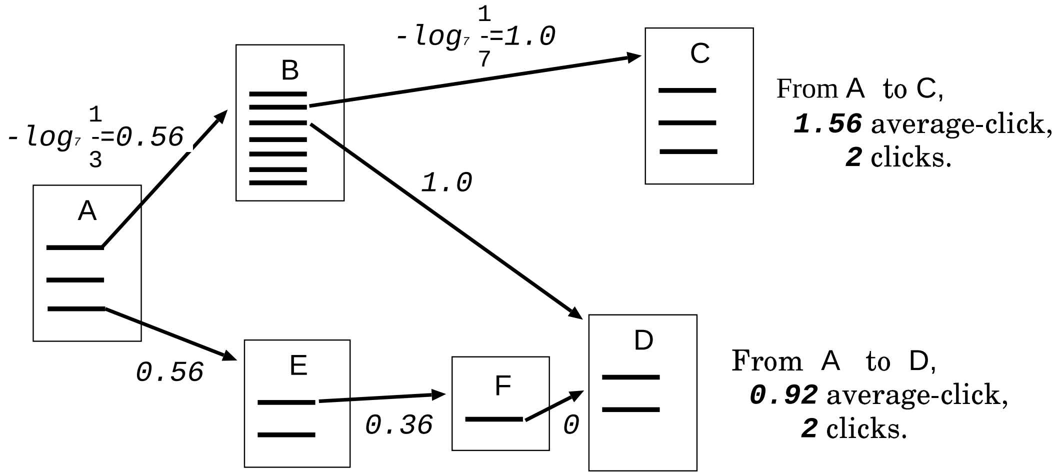

Figure 1: Average-clicks measure and clicks measure.

In the following section, we define the range of a link. Best first search is employed to find the second shortest path. The algorithm is very simple, but the results are sufficiently suggestive, shown in Section 3. We conclude the paper in Section 4.

2 Range of a Link

In this section, we detail the range of a link. First, the length of a link is defined by average-clicks measure [6]. Then, the second shortest path is defined based on the length of a link. The best first search algorithm is applied to find the second shortest path. Below we treat an *directed*, *weighted*, and *connected* graph. It means, we should proceed along a link in the right direction.

2.1 Length of a Link

Matsuo et al. proposes a new measure of distance on the Web, called *average-clicks*. It bases on the same idea as "random surfer" model as does the PageRank algorithm. The probability for a random surfer 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.

Definition 2.1
A length of a link in page p is defined as

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

We set $\alpha = 1$ for simplicity, and the base of the logarithm n to be 7, due to the fact that the average page has roughly seven hyperlinks to other pages [2].

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

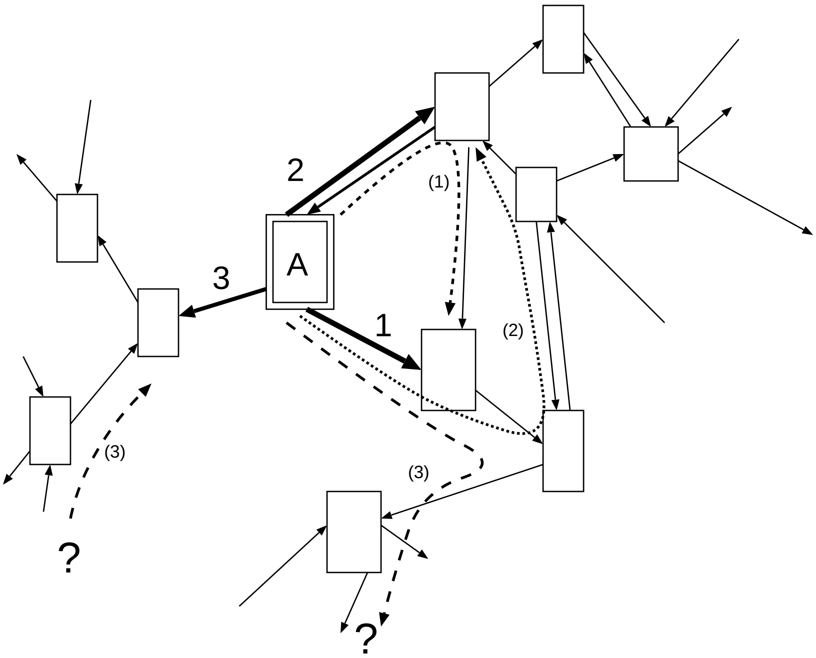

Figure 2: Second shortest path.

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.

2.2 Second Shortest Path

Next, we define the range of a link. Below the notation $p \dashrightarrow q$ denotes a link in page p to page q.

Definition 2.2
If a link $p \dashrightarrow q$ exists, the range $R(p, q)$ is defined by the length of the shortest path from p to q in the *absence* of that link. If there is no other path, $R(p, q) = \infty$.

In other words, the range $R(p, q)$ is the length of the second shortest path between p and q. The alternativeness of a link can be measured by the range. If the range of a link is large, it is a precious link. If the range is small, there can be another path around the page.

Figure 2 illustrates the second shortest path of each link in page A. The range of link 1 is the smallest among the three links because we can easily get the pointed page by another path (shown by "(1)"). Link 2 is more valuable and link 3 is the most valuable, because it points to another community.

2.3 Algorithm

To search the second shortest path from page s (stated as *start page*) to page t (stated as *target page*), the best first search is employed as Fig.3. Some modifications are made to implement the algorithm efficiently.

```
function Search_Second_Shortest_Path ($start\_page$, $target\_page$, $d_{thre}$)
  $\alpha \leftarrow 1.0$,   $n \leftarrow 7$.
  $list \leftarrow$ 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 $n_p$
      $d(p_k) \leftarrow d(p) - log(\alpha / n_p)$
      if $p$ is start_page and $p_k$ is target_page  then next
      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.

Figure 3: The best first search for the indirect shortest path.

3 Example of Hidden Relation

In this section, we show some examples of measuring the distance of the second shortest path. Since we did not have access to a large crawl of the Web, it was not feasible to do the full rank computations of links.

Table 1 is an example of links in the homepage written by one of this paper's authors. This homepage, "www.miv.t.u-tokyo.ac.jp/~matsuo" is stated below as page a. The author of page a made links to some pages, not considering consciously the meaning of the link. However, by looking this table, the author can get a lot of information he was not aware of before. The range of a link to one of the co-authors or University of Tokyo is small, because these links are common around page a. In fact, the second shortest path to the co-author includes the mutual friend's homepage.

The author of page a is greatly interested in Sumo wrestling, however, unfortunately the range of the link to Sumo wrestling official homepage is very large. Around the author, there is no such an enthusiast of Sumo wrestling. On the other hand, the range of the link to "J. League" (Japanese football league) official homepage is much nearer than Sumo wrestling. Football is more popular around the author.

And the link to a friend in a private softball team is beyond the search scope. He belongs to a different community, working for a publishing company, thus the link (or relationship) to him might be precious.

Table 1: The range of links in the homepage.

Link to	Range of the link
URL	
Second shortest path	Cumulative distance
One of the co-author's homepage	2.12
http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa	2.12
http://www.miv.t.u-tokyo.ac.jp/~matumura/research.html	1.77
http://www.miv.t.u-tokyo.ac.jp/JAICO/	1.03
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
University of Tokyo	3.78
http://www.u-tokyo.ac.jp/index.html	3.78
http://www.miv.t.u-tokyo.ac.jp/HomePageEng.html	1.63
http://www.miv.t.u-tokyo.ac.jp/~matsuo/homepageeng.html	1.03
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.00
J. League (Japanese football league) official homepage	4.02
http://www.j-league.or.jp/index.html	4.02
http://www.miv.t.u-tokyo.ac.jp/~tomobe/	2.54
http://www.miv.t.u-tokyo.ac.jp/member/present-mem.htm	1.03
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
Sumo wrestling official homepage	7.76
http://www.wnn.or.jp/wnn-t/index_e.html	7.76
http://www.ntt.co.jp/SQUARE/www-in-JP.html	4.17
http://www.ic.u-tokyo.ac.jp/index.html	2.63
http://www.u-tokyo.ac.jp/	1.03
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
One of the author's friends in a softball team	More than 8
http://www.geocities.co.jp/Athlete-Athene/6353/	

4 Conclusion

In this paper, we have shown how to rank a link by searching the second shortest path between two pages. The direct link is, so to speak, an external relation, while the second shortest path is a hidden relation. We believe that showing a hidden relation is sometimes very informative to users. Future work includes improving the algorithms, develop a system which can visually show the result, and analyze the benefit of providing users the hidden relations.

References

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

[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] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th WWW Conf.*, 1998.

[4] 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.

[5] D. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton, 1999.

[6] Y.Matsuo, Y.Ohsawa, and M.Ishizuka. Average-click: A new definition of distance on the world wide web. In *WI-2001*, 2001. to appear.