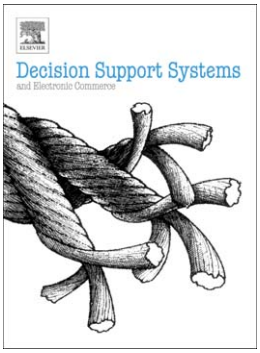

Constructing a Reliable Web Graph with Information on Browsing Behavior

Yiqun Liu, Yufei Xue, Danqing Xu, Ronwei Cen, Min Zhang, Shaoping Ma, Liyun Ru

PII: S0167-9236(12)00184-4
DOI: doi: 10.1016/j.dss.2012.06.001
Reference: DECSUP 12111

To appear in: *Decision Support Systems*

Received date: 17 March 2010
Revised date: 30 May 2012
Accepted date: 13 June 2012

Please cite this article as: Yiqun Liu, Yufei Xue, Danqing Xu, Ronwei Cen, Min Zhang, Shaoping Ma, Liyun Ru, Constructing a Reliable Web Graph with Information on Browsing Behavior, *Decision Support Systems* (2012), doi: 10.1016/j.dss.2012.06.001

Constructing a Reliable Web Graph with Information on Browsing Behavior

Yiqun Liu[1], Yufei Xue, Danqing Xu, Ronwei Cen, Min Zhang, Shaoping Ma, Liyun Ru

State Key Lab of Intelligent Technology & Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University

[1] Corresponding author. Contact Information: FIT Building 1-506, Tsinghua University, Beijing, 100084, China P.R., Tel.: +86-10-62796672, Fax: +86-10-62796672, E-mail: yiqunliu@tsinghua.edu.cn.

Abstract

Page quality estimation is one of the greatest challenges for Web search engines. Hyperlink analysis algorithms such as PageRank and TrustRank are usually adopted for this task. However, low quality, unreliable and even spam data in the Web hyperlink graph makes it increasingly difficult to estimate page quality effectively. Analyzing large-scale user browsing behavior logs, we found that a more reliable Web graph can be constructed by incorporating browsing behavior information. The experimental results show that hyperlink graphs constructed with the proposed methods are much smaller in size than the original graph. In addition, algorithms based on the proposed "surfing with prior knowledge" model obtain better estimation results with these graphs for both high quality page and spam page identification tasks. Hyperlink graphs constructed with the proposed methods evaluate Web page quality more precisely and with less computational effort.

HIGHLIGHTS

1. With user browsing behavior information, it is possible to improve the performance of quality estimation results for commercial search engines.
2. Three different kinds of Web graphs were proposed which combines original hyperlink and user browsing behavior information.
3. Differences between the constructed graphs and the original Web graph show that the constructed graphs provide more reliable information and can be adopted for practical quality estimation tasks.
4. The incorporation of user browsing information is more important than the selection of link analysis algorithms for the task of quality estimation.

Keywords

Web graph; Quality estimation; Hyperlink analysis; User behavior analysis; PageRank

1. Introduction

The explosive growth of data on the Web makes information management and retrieval increasingly difficult. For contemporary search engines, estimating page quality plays an important role in crawling, indexing and ranking processes. For this reason, the estimation of Web page quality is considered as one of the greatest challenges for Web search engines [15].

Currently, the estimation of page quality mainly relies on an analysis of the hyperlink structure of the Web. The success of PageRank [25] and other hyperlink analysis algorithms such as HITS (Hyperlink-Induced Topic Search) [19] and TrustRank [11] shows that it is possible to estimate Web page quality query independently. These hyperlink analysis algorithms are based on two basic assumptions [8]: First, if two pages are connected by a hyperlink, the page linked is recommended by the page that links to it (recommendation). Second, the two pages share a similar topic (locality). Hyperlink analysis algorithms adopted by both commercial search engines (such as [5, 12, 21, 25]) and researchers (such as [11, 13, 14, 19, 20]) all rely on these two assumptions. However, these two assumptions miss subtleties in the structure of the actual Web graph. The assumptions and the consequent algorithms thus face challenges in the current Web environment.

For example, Table 1 shows several top Web sites ranked by PageRank on a Chinese Web corpus[2] of over 130 million pages. To determine whether the PageRank score accurately represents the popularity of a Web site, we also gathered traffic rankings as measured by Alexa.com.

[2] The Corpus is called SogouT corpus. It contains 130 million Chinese Web pages and was constructed in July 2008. Web site: http://www.sogou.com/labs/dl/t.html

Table 1. Top-ranked Web sites by PageRank in a Chinese Web hyperlink graph

Web Site	Ranked by PageRank	Ranked by Alexa.com[3] traffic rankings in China
www.hd315.gov.cn	2	1,655
www.qq.com	3	*2*
www.baidu.com	6	*1*
www.miibeian.gov.cn	7	179
www.sina.com.cn	9	*3*

The data in Table 1 show that several of the top 10 Web sites as ranked by PageRank also received a large number of user visits. For example, www.baidu.com, www.qq.com and www.sina.com.cn are also the three most frequently visited Web sites in China according to Alexa.com (their traffic rankings are shown in Table 1 in italics). In contrast, several top-ranked sites received a relatively small number of user visits, such as www.hd315.gov.cn and www.miibeian.gov.cn. According to [25], pages with high PageRank values are either well cited from many places around the Web or pointed to by other high PageRank pages. In either case, the pages with the highest PageRank values should be frequently visited by Web users because PageRank can be regarded as "the probability that a random surfer visits a page". Traffic is also considered as one of the possible applications of PageRank algorithm in [25]. However, these top-ranked sites do not receive as many user visits as their PageRank rankings indicate. Although authority does not necessarily mean high traffic on the Web, we believe that either the MII site or the www.hd315.gov.cn site should not be ranked so high in quality estimation results because there are many other government agencies which are also authoritative but ranked much lower than these two sites.

In order to find out why the MII site and the www.hd315.gov.cn are ranked so high according to PageRank score, we examine the hyperlink structure of these sites. Figure 1 shows how

[3] http://www.alexa.com/topsites/countries/CN

www.baidu.com (the most popular Chinese search engine) links to www.miibeian.gov.cn (home page of the Ministry of Industry and Information Technology of China). As shown in the red box, the hyperlink is located at the bottom of the page, and the anchor text contains the Web site's registration information. Each Web site in China should register to the Ministry of Industry and Information Technology (MII), and site owners are requested to put the registration information on each page. Therefore, almost all Web sites in China link to the MII Web site, and the PageRank score of www.miibeian.gov.cn is very high because of the huge number of in-links. The Web site www.hd315.gov.cn is highly ranked by PageRank for a similar reason; each commercial site in China is required to put registration information on their pages, and the registration information contains a hyperlink to www.hd315.gov.cn.

Figure 1. A sample site (http://www.baidu.com) which links to www.miibeian.gov.cn, the site in the sample corpus with the 7th highest PageRank score.

From this example we can see that quality estimation results given by PageRank on practical Web environment may not be so reasonable. Web sites such as the MII site are ranked quite high because many Web pages link to them. However, many of these hyperlinks are created due to legal, commercialized or even spamming reasons. Hyperlinks on Web graph should not be

treated as equally important as PageRank supposes in [2]. Practical Web users do not act like the "random surfer"; instead, they only click hyperlinks interesting to them. Therefore, Web sites that are connected by hyperlinks that Web users are not interested in clicking usually get high PageRank score which they do not deserve.

This example shows that hyperlink analysis algorithms are not always successful in the real Web environment because of the existence of hyperlinks that users seldom click. Removing these hyperlinks from Web graph is an important step in constructing a more reliable graph on which link analysis algorithms can be performed more effectively.

To reduce noises in the Web graph, we analyze information on users' browsing behaviors collected by search engine toolbars or browser plug-in software. Information on browsing behavior can reveal which pages or hyperlinks are frequently visited by users and which are not, allowing construction of a more reliable Web graph. For example, although many pages link to the MII homepage, few people click on these links because site registration information is not interesting to most Web users. These hyperlinks may be regarded as "meaningless" or "invalid" because they are not involved in users' Web surfing process. If we construct a new Web graph without these links, the representation of users' browsing behavior will not be affected, but the PageRank score calculated by the new graph will be more accurate because most of the hyperlinks connecting to the MII homepage are removed.

The number of users visiting a site can be regarded as implicit feedback about the importance of both hyperlinks and pages in the Web graph. However, constructing a more reliable graph with this kind of information remains a challenging problem. Retaining only the nodes and vertexes that have been visited at least once is one potential option. Several researchers, such as Liu *et al.*

[23], have constructed such a graph, called a 'user browsing graph', and have used it to gain better estimates of page quality than with the original Web graph[4]. However, with user browsing information, there are other options in constructing a Web graph other than the user browsing graph. The contributions of our work include:

- With user browsing information, a new Web surfing model is constructed other than the "random surfer model" adopted by previous researches such as PageRank. This "surf with prior knowledge model" incorporates both user behavior information and hyperlink information and is a better simulation of Web users' surfing processes.
- Two quality estimation algorithms (userPageRank and userTrustRank) are proposed according to the new "surf with prior knowledge model". These algorithms take user preference of hyperlinks into consideration and they can be performed on the user browsing graph.
- Two different kinds of Web graph construction algorithms are proposed besides user browsing graph to combine both browsing and hyperlink structure information. Characteristics and evolution of these graphs are studied and compared with the original Web graph.

The remainder of the paper is organized as follows: Section 2 gives a review of related work on page quality estimation and user browsing behavior analysis. Section 3 introduces the "surf with prior knowledge model" and the quality estimation algorithms based on it. Section 4 presents algorithms for constructing Web graphs based on both user browsing and hyperlink information. Section 5 describes the structure and evolution of the Web graphs constructed with the proposed algorithms. The experimental results of applying different algorithms to estimate page quality on different graphs are reported in Section 6. Conclusions and future work are provided in

[4] Original Web graph is the Web graph constructed with pages and hyperlinks collected from the real Web environment without removing noises.

Section 7.

2. Related Work

2.1 Page Quality Estimation

Most previous work on page quality estimation focuses on exploiting the hyperlink graph of the Web and builds a model based on that graph. Since the success of PageRank **Error! Reference source not found.** in the late 1990s, extensive research has attempted to improve the efficiency and effectiveness of the original algorithm [12, 13, 14]. However, the basic idea has not changed: a Web page's quality is evaluated by estimating the probability of a Web surfer's visiting the page using a random walk model. The HITS algorithm evaluates Web page quality using two different metrics, the hub score and authority score. Experimental results based on both IBM CLEVER search system evaluation **Error! Reference source not found.** and human experts' annotations [1] have demonstrated the effectiveness of HITS.

In addition to methods to evaluate the quality of Web pages, researchers have proposed link analysis algorithms to identify spam pages. Spam pages are created with the intention of misleading search engines. Gyongyi *et al.* [11] developed the TrustRank algorithm to separate reputable pages from spam. This work was followed by other methods based on the link structure of spam pages, such as Anti-Trust Rank [20] and Truncated PageRank [2] algorithms. TrustRank is an effective link analysis algorithm that assigns a trust score to Web pages. Pages with low trust scores tend to be spam pages, and pages with high trust scores tend to be high quality pages.

These link analysis algorithms have become popular and important tools in search engines' ranking mechanisms. However, the Web graph on which these algorithms are based is not particularly reliable because hyperlinks can be easily added or deleted by page authors or even

by Web users (via Web 2.0 services). Therefore, as shown in Table 1, noise in Web graphs makes it difficult for these algorithms to evaluate page quality effectively.

Several methods have been proposed to counteract the manipulation of Web structure. Algorithms such as DiffusionRank [28] and AIR (Affinity Index Ranking) [18] were designed to fix the flaws of PageRank and TrustRank. DiffusionRank is motivated by the phenomenon of heat diffusion, which is analogous to the dissipation of energy via out-links. AIR scores for Web pages are obtained by using an equivalent electronic circuit model. Similar to TrustRank, both algorithms require the construction of a "high quality seed set". Experimental results have shown that DiffusionRank and AIR perform better than PageRank and TrustRank in removing spam both on toy graphs and in real Web graphs. However, aside from hyperlinks generated for Web structure manipulation and spam, most Web pages contain meaningless and low quality hyperlinks such as copyright links, advertisement links, and registration information links and so on. These links are not popular and are seldom clicked by users, but they comprise a large part of Web graphs. Both DiffusionRank and AIR algorithms are unable to deal with this kind of "noise" in hyperlink structure data.

Because of the problems that hyperlink analysis algorithms encounter in real Web environment, researchers have tried to use features other than hyperlinks to evaluate quality of Web pages. Chau et al. [7] have identified pages on certain topics using both content-based and link-based features. Liu *et al.* [24] have proposed a learning-based method for identifying search target pages query independently using content-based and hyperlink-based features, such as document length and in-link count. Jacob *et al.* [16] have also adopted both content-based and hyperlink-based approaches to detect Web spam. Although these methods use features other

than links, link analysis algorithms still play an important role in the identification of high quality pages or spam pages. Therefore, the quality of Web hyperlink data and the effectiveness of link analysis algorithms remain challenging problems.

In contrast to these approaches, we incorporate Web users' browsing behavior to indicate page quality. Most users' browsing behavior is driven by their interests and information needs. Therefore, pages that are visited and hyperlinks that are clicked by users should be regarded as more meaningful and more important than those that are not. It is therefore reasonable to use users' preferences to prune the hyperlink graph.

2.2 User Browsing Behavior Analysis

Although researchers such as Page et. al. [25] tried to incorporate browsing information (collected from DNS providers) in page quality estimation at the early stage of hyperlink analysis researches, browsing behavior analysis has not become popular until recent years. Web browser toolbars such as Google Toolbar and Live Toolbar collect user browsing information. It is considered as an important source of implicit feedback on page relevance and importance and was widely adopted in Web site usability [10, 17, 26], user intent understanding [27] and Web search [4, 22, 23, 29] researches.

Using this information on browsing behavior, it is possible to prune the Web graph by removing unvisited nodes and links. For example, Liu *et al.* [23] constructed a "user browsing graph" with Web access log data. It is believed that the user browsing graph can avoid most of the problems of the original Web graph because links in the browsing graph are actually chosen and clicked by users. Liu *et al.* also proposed an algorithm to estimate page quality, BrowseRank, which is based on continuous-time Markov process model. Their study shows that the BrowseRank algorithm works better than hyperlink analysis algorithms such as PageRank and TrustRank

when the latter two algorithms are performed on the whole Web graph.

The user browsing graph is not the only way to incorporate browsing behavior into page quality estimation. In addition, the interpretation of the user browsing graph is not obvious. For example, we can infer that the user browsing graph differs from the whole Web graph in some aspects, but precisely how do the structures of these two graphs differ from each other? How does the user browsing graph evolve over time? BrowseRank outperforms PageRank and TrustRank algorithms when the latter two algorithms are performed on the original Web graph, but how do hyperlink analysis algorithms perform on the user browsing graph?

We try to answer these questions through experimental studies, and we also attempt to determine how data on users' browsing behavior can be better analyzed to construct a more reasonable Web surfing model rather than the widely adopted random surfer model.

3. Surfing with Prior Knowledge

With the example shown in Table 1 and Figure 1, we know that hyperlinks are not clicked by users with equal probabilities and they should not be treated as equally important in the construction of surfing models. However, due to the difficulties in collecting user browsing information, most previous works on Web graph mining are based on the "random surfer model" which supposes user simply keeps clicking on successive links at random.

Differently from these works, we collected a large amount of user browsing information with the help of a widely used search engine in China. These Web-access logs were collected from Aug. 3, 2008, to Oct. 6, 2008 (60 days; logs from Sept. 3 to Sept. 7 were not included because of hard disk failure). Over 2.8 billion hyperlink click events were recorded and can be adopted as prior knowledge in the construction of surfing models. Details of these log data are

introduced in Section 4.1.

Designed with random surfer model, one of the major flaws of the PageRank algorithm is "over-democracy" [28]. The original algorithm assumes that the Web user either randomly follows a hyperlink on a Web page and navigates to the destination (with probability α) or randomly chooses a different page on the given Web graph (with probability 1-α).

$$PageRank^{(k+1)}(X) = \alpha \cdot \sum_{X_i \Rightarrow X} \frac{PageRank^{(k)}(X_i)}{\#Outlink(X_i)} + (1-\alpha) \cdot \frac{1}{N} \qquad \textbf{(1)}$$

According to Equation (1), the PageRank score of a page is divided evenly between all of its outgoing hyperlinks. However, hyperlinks on Web pages are not equally important. Some hyperlinks, such as "top stories" links on the CNN.com homepage, are more important, whereas others, such as advertisements, are less important.

Therefore, it is not reasonable to assume that users will follow hyperlinks on a Web page with equal probabilities. If we introduce the probability of visiting page X_j directly after visiting page X_i, namely $P(X_i \Rightarrow X_j)$, the random surfer model will be replaced by the "*surfing with prior knowledge*" model and the estimation of $P(X_i \Rightarrow X_j)$ requires prior knowledge of user browsing behaviors.

With the "*surfing with prior knowledge*" model, Web users do not click on hyperlinks on the Web pages they are visiting randomly, instead, each hyperlink L is clicked with a probability of $P(X_i \Rightarrow X_j)$ in which X_i is the source page and X_j is the destination page of the L.

With the new surfing model, Equation 1 can be modified as follows:

$$PageRank^{(k+1)}(X) = \alpha \cdot \sum_{X_i \Rightarrow X} PageRank^{(k)}(X_i) P(X_i \Rightarrow X) + (1-\alpha) \cdot \frac{1}{N} \qquad \textbf{(2)}$$

In Equation (2), $P(X_i \Rightarrow X_j)$ is the probability of visiting page X directly after visiting page

X_i. However, for the original Web graph, it is not possible to estimate this probability because the relevant information is not provided. Therefore, PageRank (as well as TrustRank) has to be computed using equal $P(X_i \Rightarrow X_j)$ values (as Equation (1)).

To incorporate prior user browsing information into the original Web graph, the user-visited nodes and edges should be selected and the number of user clicks on each hyperlinks (edges) should be recorded. With this information, we can decide which hyperlinks are important and estimate the probability of $P(X_i \Rightarrow X_j)$ with the maximum likelihood assumption.

If we use $UC(X_i \Rightarrow X_j)$ to represent the number of user clicks from X_i to X_j, the original PageRank algorithm can be modified as follows:

$$userPageRank^{(k+1)}(X) = \alpha \cdot \sum_{X_i \Rightarrow X} userPageRank^{(k)}(X_i) \frac{\#UC(X_i \Rightarrow X)}{\sum_{X_i \Rightarrow X_j} \#UC(X_i \Rightarrow X_j)} + (1-\alpha) \cdot \frac{1}{N} \quad \textbf{(3)}$$

In Equation 3, the probability of $P(X_i \Rightarrow X_j)$ is estimated by the weighted *UC* factor with maximum likelihood assumption. The PageRank of page X_i is divided between the outgoing links, weighted by *UC* of each link. Aside from this PageRank division, no other part of the original algorithm is changed. Therefore, the time complexity and the efficiency of this algorithm stay the same.

A similar modification can be applied to the TrustRank algorithm, which traditionally divides the trust score equally between outgoing links. The original and the modified algorithms are shown in Equations (4) and (5) separately.

$$TrustRank^{(k+1)}(X) = \alpha \cdot \sum_{X_i \Rightarrow X} \frac{TrustRank^{(k)}(X_i)}{\#Outlink(X_i)} + (1-\alpha) \cdot \mathrm{d} \quad \textbf{(4)}$$

$$userTrustRank^{(k+1)}(X) = \alpha \cdot \sum_{X_i \Rightarrow X} userTrustRank^{(k)}(X_i) \frac{\#UC(X_i \Rightarrow X)}{\sum_{X_i \Rightarrow X_j} \#UC(X_i \Rightarrow X_j)} + (1-\alpha) \cdot \mathrm{d} \quad \textbf{(5)}$$

With the "*surfing with prior knowledge*" model, hyperlinks on Web pages are not treated as equally important, instead, the probability of user clicking are estimated with prior knowledge and maximum likelihood assumption. By this means, we hope to improve the performance of PageRank and TrustRank which are originally based on the random surfer model.

We believe that the new surfing model can also be utilized to other graphs besides the Web hyperlink graph if the probability of visiting one node from another can be estimated. For example, let *G*=(*V*, *E*) denotes a social graph, where *V* represents the users and *E* represents the relationship between them. In many Web-based social network services such as twitter and weibo[5], the relationship between users can be described as a directed edge from follower to followee, which is similar to the hyperlink from source page to destination page.

Intuitively, the influence of s social node in social networks is similar to the quality score of a Web page. It means that if we try to estimate influence scores on a social graph, hyperlink algorithms such as PageRank and TrustRank can also be utilized. As hyperlinks in a Web graph, we believe that the "following" relationships between nodes in a social graph are also not equally important. This is because users may follow another user for different reasons and closest relationships should be valued more. Therefore, "*surfing with prior knowledge*" model is also more reasonable than the random surfer model on the social graph although the prior knowledge ($P(X_i \Rightarrow X_j)$) should be estimated by a different means.

4. Web Graph Construction with Information on Browsing Behavior

4.1 Data on User Browsing Behavior

Based on the "*surfing with prior knowledge*" model described in Section 3, we revise the original PageRank and TrustRank algorithm by incorporating prior user browsing behavior

[5] Weibo (http://www.weibo.com) is China's largest microblog service provider which owns over 250 million users.

information. Therefore, the newly proposed userPageRank and userTrustRank algorithms require additional information and cannot be performed on the original Web graph. To construct a reliable Web graph that incorporates user browsing behavior information, we collected data on users' browsing behavior (also called Web-access log data or Web usage data). In contrast to log data from search engine queries and click-through data, this kind of data is collected using browser toolbars. It contains information on Web users' total browsing behavior, including their interactions with search engines and other Web sites.

To provide value-added services to users, most browser toolbars also collect anonymous click-through information on users' browsing behavior. Previous work such as [4] has used this kind of click-through information to improve ranking performance. Liu *et al.* [22] have proposed a Web spam identification algorithm based on this kind of user behavior data. In this paper, we also adopt Web access logs collected by toolbars because this enables us to freely collect users' browsing behavior information with no interruption to the users. An example of the information recorded in these logs is shown in Table 2 and Example 1.

Table 2. Information recorded in Web-access logs

Name	Description
Time Stamp	Date/Time of the click event
Session ID	A randomly assigned ID for each user session
Source URL	URL of the page that the user is visiting
Destination URL	URL of the page to which the user navigates

Example 1. A sample Web-access log collected on Dec. 15, 2008

(01:07:09) (3ffd50dc34fcd7409100101c63e9245b) (http://v.youku.com/v_playlist/f1707968o1p7.html) (http://www.youku.com/playlist_show/id_1707968.html)

(01:07:09) (f0ac3a4a87d1a24b9c1aa328120366b0) (http://user.qzone.qq.com/234866837) (http://cnc.imgcache.qq.com/qzone/blog/tmygb_static.htm)

(01:07:09) (3fb5ae2833252541b9ccd9820bad30f6) (http://www.qzone8.net/hack/45665.html) (http://www.qzone8.net/hack/)

Table 2 and Example 1 show that no private information was included in the log data. The

information shown can be easily recorded using browser toolbars by commercial search engine systems. Therefore, collecting this kind of information for the construction of hyperlink graphs is practical and feasible.

4.2 Construction of a User Browsing Graph and a User-oriented Hyperlink Graph

With the data on users' browsing behavior described in Section 4.1, we identified which pages and hyperlinks were visited and the following two algorithms are adopted to construct the user browsing graph and the user-oriented hyperlink graph, respectively.

Algorithm 1 constructs a graph completely based on user behavior data. Only nodes and hyperlinks that were visited at least once are added to the graph. This graph is similar to the graph constructed by Liu et al. in [23], except that the number of user visits on each edge is also recorded to estimate $P(X_i \Rightarrow X_j)$ for userPageRank and userTrustRank. Following their convention, we also call this graph user browsing graph (*BG*(*V*,*E*) for short).

1. $V = \{\}, \ E = \{\}$
2. For each record in the Web-access log, if the source URL is *A* and the destination URL is *B*, then

$if \ A \notin V, V = V \cup \{A\};$
$if \ B \notin V, V = V \cup \{B\};$
$if \ (A,B) \notin E$
$\quad E = E \cup \{(A,B)\}$
$\quad Count(A,B) = 1;$
$else$
$\quad Count(A,B)++;$

Algorithm 1. Algorithm to construct the user browsing graph.

Algorithm 2 constructs a graph distinct from *BG*(*V*,*E*). These two graphs share a common set of nodes, though the graph constructed with Algorithm 2 retains all of the edges between these nodes from the original Web graph. We call this graph a user-oriented hyperlink graph

(*user-HG*(*V,E*) for short) because it is extracted from the original Web graph but has nodes selected with user information. The original Web graph was constructed by the same search engine company that provided Web access logs to us. Collected in July 2008, it contains over 3 billion pages from 111 million Web sites and covers a major proportion of Chinese Web pages at that time.

1. $V = \{\},\ E = \{\}$
2. For each record in the Web-access log, if the source URL is *A* and the destination URL is *B*, then

   $if\ \ A \notin V, V = V \cup \{A\};$

   $if\ \ B \notin V, V = V \cup \{B\};$
3. For each *A* and each *B* in *V*,

   $if\ ((A,B) \in Original\ \ Web\ \ Graph)\ AND\ ((A,B) \notin E)$

   $E = E \cup \{(A,B)\}$

Algorithm 2. Algorithm to construct the user-oriented hyperlink graph.

Thus, both *BG*(*V,E*) and *user-HG*(*V,E*) are constructed with the help of browsing behavior data. The latter graph contains more hyperlinks, whereas the former graph only retains hyperlinks that are actually followed by users. We can see that userPageRank and userTrustRank cannot be performed on *user-HG*(*V,E*) because browsing information are not recorded for all edges on this graph.

4.3 Comparison of the User Browsing and User-Oriented Hyperlink Graphs

We constructed *BG*(*V,E*) and *user-HG*(*V,E*) with the data on user behavior described in Section 4.1. Table 3 shows how the compositions of these two graphs differ from each other.

Table 3. Differences between *BG*(*V,E*) and *user-HG*(*V,E*) in the edge sets

	#(Common edges)	#(Total edges)	Percentage of common edges
BG(V,E)	2,591,716	10,564,205	24.53%
User-HG(V,E)		139,125,250	1.86%

According to Table 3, we found that although the hyperlink graph *user-HG*(*V,E*) shares a

common set of nodes with *BG(V,E)*, the compositions of these two graphs differ significantly. First, *BG(V,E)* is less than one-tenth the size of *user-HG(V,E)*. The percentage of common pages in *user-HG(V,E)* is only 1.86%; thus, most (98.14%) of the links in *user-HG(V,E)* are not actually clicked by users. This difference is consistent with people's Web browsing experience that pages usually provide too many hyperlinks for users to click.

Another interesting finding is that the *user-HG(V,E)* graph does not include all the edges in *BG(V,E)*. Less than one-quarter of the pages in *BG(V,E)* also appear in *user-HG(V,E)*. This phenomenon can be partially explained by the fact that *User-HG(V,E)* is constructed with information collected by Web crawlers, and it is not possible for any crawler to collect the hyperlink graph of the whole Web; it is too huge and changing so fast. When we examined the links that only appear in *BG(V,E)*, we found another reason why *user-HG(V,E)* does not include them. A large proportion of these links come from users' clicks on search engines result pages (SERPs). Table 4 shows the number of SERP-oriented hyperlinks in *BG(V,E)*.

Table 4. Number of SERP-oriented edges that are not included in *user-HG(V,E)*

Search engine	**Number of edges that are not included in *user-HG(V,E)***
Baidu.com	1,518,109
Google.cn	1,169,647
Sogou.com	291,829
Soso.com	147,034
Yahoo.com	143,860
Total	*3,270,479 (30.96% of all edges in BG(V,E))*

Tables 3 and 4 reveal that of the links that appear only in *BG(V,E)* (7.97 million edges in total), over 3.27 million come from SERPs of the five most frequently used Chinese search engines. This number constitutes 30.96% of all edges in *BG(V,E)*. Web users click many links on SERPs, but almost none of these links would be collected by crawlers. These links contain valuable

information because they link to Web pages that are both recommended by search engines and clicked by users. It is not possible for Web crawlers to collect all of the links from SERPs without information on user behavior because the number of such links would be overwhelmingly large.

Another important type of links that appear only in *BG(V,E)* are hyperlinks that are clicked in users' password-protected sessions. For example, login authorization is sometimes needed to visit blog pages. After logging in, Web users often navigate among these pages, and Web-access logs can record these browsing behaviors. However, ordinary Web crawlers cannot collect these links because they are not allowed to access the contents of protected Web pages.

4.4 Construction of the User-oriented Combined Graph

Section 4.3 shows that the user browsing graph differs from the user-oriented hyperlink graph in at least two ways: First, compared with *user-HG(V,E)*, a large fraction of the edges (98.14% of *E* in *user-HG(V,E)*) are omitted from *BG(V,E)* because they are not clicked by any user. Second, *BG(V,E)* contains hyperlinks that are difficult or impossible for Web crawlers to collect. Thus, each graph contains unique information that is not contained by the other graph. Therefore, if we construct a graph containing all of the hyperlinks and nodes in *BG(V,E)* and *user-HG(V,E)*, it should contain more complete hyperlink information. We adopt the following algorithm (Algorithm 3) to construct such a graph, which combines all of the hyperlink information in *BG(V,E)* and *user-HG(V,E)*.

1. $V = \{\}, \ E = \{\}$
2. For each record in the Web-access log, if the source URL is *A* and the destination URL is *B*, then

   $if \ \ A \notin V, V = V \cup \{A\};$

   $if \ \ B \notin V, V = V \cup \{B\};$

3. For each A and each B in V,

$$if\ ((A,B) \in BG(V,E)) OR ((A,B) \in userHG(V,E))$$
$$E = E \cup \{(A,B)\}$$

Algorithm 3. Algorithm to construct the user-oriented combined graph.

This algorithm can construct a graph that shares the same node set as *BG(V,E)* and *user-HG(V,E)* but that contains the hyperlinks of both graphs. Because it combines the edge sets of *BG(V,E)* and *user-HG(V,E)*, we call it a user-oriented combined graph (*user-CG(V,E)* for short). Similar with *user-HG(V,E)*, it doesn't contain clicking information on all the edges and userPageRank/userTrustRank cannot be performed on it.

4.5 Stats of the Constructed Graphs

With the data from Web-access logs described in Section 4.1 and the original whole Web graph (named *whole-HG(V,E)* for short) mentioned in Section 4.2, we constructed three graphs (*BG(V,E)*, *user-HG(V,E)*, and *user-CG(V,E)*). These graphs were constructed at the site-level instead of the page-level to improve efficiency. This level of resolution is also appropriate because a large number of search engines adopt site-level link analysis algorithms and then obtain page-level link analysis scores using a propagation process within Web sites. Another problem with a page-level graph is that due to data sparsity problem, there are only a few user visits for a large part of pages and the behavior data may be not so reliable. However, for a site-level graph, the average number of user visits per site is much larger and data sparsity can be avoided to a large extent. According to experimental results in our previous work [29], we also found that a site-level model outperformed a page-level model because the average number of browsing activities per site is much larger, indicating more reliable behavior information sources.

Descriptive statistics of these constructed graphs are shown in Table 5.

Table 5. Sizes of the constructed and the original Web graphs

Graph	Vertices (#)	Edges (#)	Edges/Vertices
BG(V,E)	4,252,495	10,564,205	2.48
user-HG(V,E)	4,252,495	139,125,250	32.72
user-CG(V,E)	4,252,495	147,097,739	34.59
whole-HG(V,E)	110,960,971	1,706,085,215	15.38

We can see from Table 5 that *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)* cover a small percentage (3.83%) of the vertices of the original Web graph. The edge sets of these three graphs are also much smaller than the Web graph, but the average number of hyperlinks per node in *user-HG(V,E)* and *user-CG(V,E)* is higher than that of *whole-HG(V,E)*. This result means that user-accessed nodes are more strongly connected to each other than the other parts of the Web. This pattern hints the presence of a large SCC (Strongly Connected Component) proposed in [9] in the user browsing graphs. Another finding is that compared with *user-HG(V,E)* and *user-CG(V,E)*, the ratio of edges to vertices in *BG(V,E)* is much smaller. Thus, a large fraction of hyperlinks are removed for this graph because they are not followed by users. The retained links are ostensibly more reliable than the others, however; whether this information loss creates problems for link analysis algorithms remains to be determined.

5. Structure and Evolution of Constructed Graphs

5.1 Structure of the Constructed Graphs

The degree distribution has been used to describe the structure of the Web by many researchers, such as Broder et al. [6]. The existence of a power law in the degree distribution has been verified by several Web crawls [6, 9] and is regarded as a basic property of the Web. We were interested in whether power laws could also describe the in-degree and out-degree distributions in the constructed graphs. Experimental results of degree distributions of both *BG(V,E)* and *user-HG(V,E)* are shown in Figures 2 and 3. We did not consider the degree distributions of

user-CG(V,E) because it is a combination of *BG(V,E)* and *user-HG(V,E)*. If in-degree and out-degree distributions of these two graphs follow a power law, *user-CG(V,E)* will as well.

Figure 2 shows that in-degree distributions of both *BG(V,E)* and *user-HG(V,E)* follow a power law. The exponent of the power law (1.75) is smaller than that found in previous results (approximately 2.1 in [6, 9]). This difference is because our hyperlink graph is based on sites, whereas previous graphs were based on pages. There are fewer unpopular (low in-degree) nodes in a site-level graph compared with a page-level graph because a large number of unpopular pages may come from the same Web site. Another phenomenon is that the exponent of power law distribution in *BG(V,E)* (2.30) is larger than that of *user-HG(V,E)* (1.75). This differences implies that with an increase in in-degree *i*, the number of vertices with *i* in-links drops faster in the user browsing graph. This pattern can be explained by the fact that some Web sites are relatively more popular (have higher in-degree) in the user browsing graph than in the user-oriented hyperlink graph.

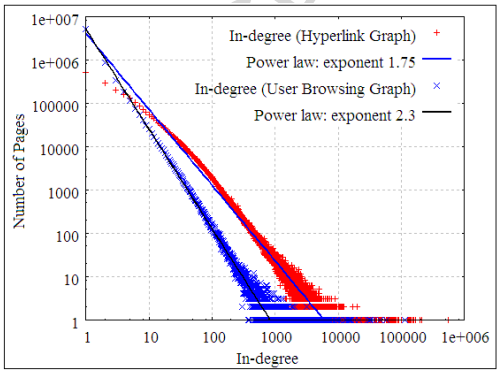

Figure 2. In-degree distributions of both *BG(V,E)* and *user-HG(V,E)* subscribe to the power law.

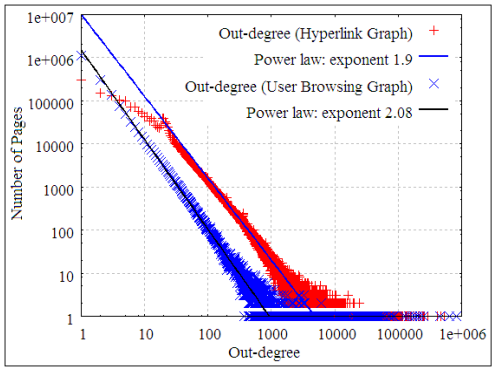

Figure 3. Out-degree distributions of both *BG(V,E)* and *user-HG(V,E)* subscribe to the power law.

The out-degree distributions of both graphs also subscribe to the power law (Figure 3). The exponent of the out-degree distribution in a page-based graph has been estimated to be 2.7 [6, 9]. The exponent estimated for our site-based graph is much smaller (1.9). In a site-based graph,

out-links that link to pages in the same site are omitted. This assumption reduces the number of out-links of many vertices and reduces the difference between high and low out-link vertices. The exponent of the out-degree distribution in *BG(V,E)* is larger than the one in *user-HG(V,E)*. As for the out-degree distribution, this differences means with the increase in out-degree *o*, the number of vertices with *o* in-links drops faster in the user browsing graph.

The experimental results shown in Figures 2 and 3 confirm that similar to the whole Web graph, the in-degree and out-degree distributions of both *BG(V,E)* and *user-HG(V,E)* follow a power law. However, the exponents of the power law distributions are different because the constructions of *BG(V,E)* and *user-HG(V,E)* decrease the numbers of valueless nodes and hyperlinks compared with the original Web graph. The fact that *BG(V,E)* and *user-HG(V,E)* inherit characteristics of the whole Web makes it possible for us to perform state-of-the-art link analysis algorithms on these graphs.

5.2 Evolution of *BG(V,E)* and Quality Estimation of Newly visited Pages

The purpose of our work is to estimate Web page quality with the help of information on user browsing behavior. For practical search engine applications, an important issue is whether the page quality scores calculated off-line can be adopted for on-line search process. *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)* were all constructed with browsing behavior information collected by search engines. This kind of information is collected during a certain time period. Therefore, user behavior outside this time period cannot be included in the construction of these graphs. If pages needed by users are not included in the graphs, it is impossible to calculate their quality scores. Therefore, it is important to determine how the compositions of these graphs evolve over time and whether newly visited pages can be included in the graphs.

To determine whether the construction of *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)* can avoid

the problem of newly visited and missing pages, we designed the following experiment:

Step 1. A large number of pages appear each day, and only a fraction of them are visited by users. We only focus on the newly visited pages that are actually visited by users because the absence of pages from the graph could affect users' browsing experiences. Therefore, we examine how many newly visited pages are included by the constructed graphs.

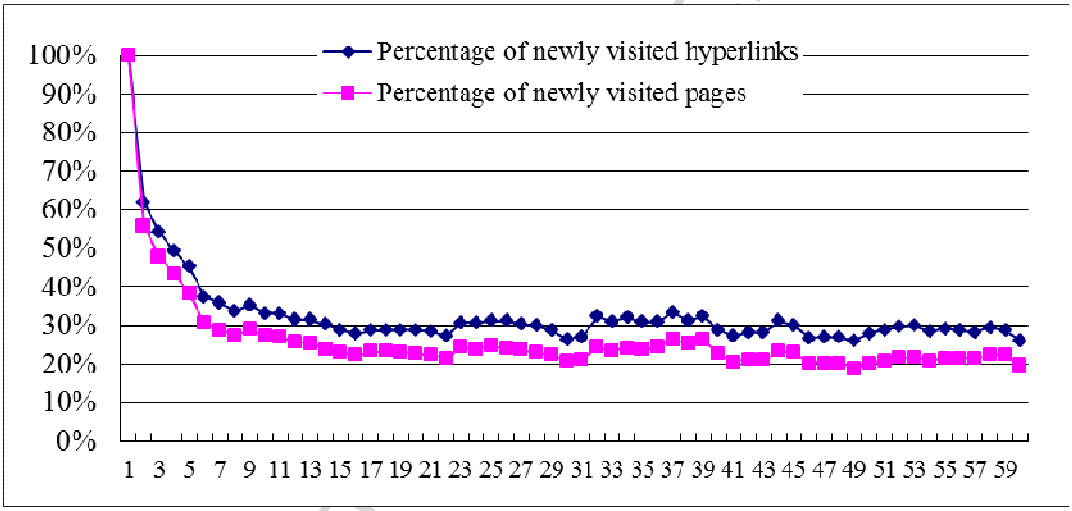

Figure 4. Evolution of *BG(V,E)*. Category axis: day number, assuming Aug. 3, 2008, is the first day. Value axis: percentage of newly clicked pages/hyperlinks not included in *BG(V,E)* (*BG(V,E)* is constructed with data collected from the first day to the given day).

In Figure 4, each data point shows the percentage of newly clicked pages/hyperlinks that are not included by *BG(V,E)*. On each day, *BG(V,E)* is constructed with browsing behavior data collected from Aug. 3, 2008, (the first day in the figure) to the date before that day. We focus on *BG(V,E)* because *user-HG(V,E)* and *user-CG(V,E)* share the same vertex set. On the first day, all of the edges and vertices are newly visited because no data has yet been included in *BG(V,E)*. From the second day to approximately the 15th day, the percentage of newly visited edges and vertices drops. On each day after the 15th day, approximately 30% of the edges and 20% of the vertices are new to the *BG(V,E)*, which is constructed with data collected before that day.

During the first 15 days, the percentage of newly visited edges and vertices drops because the

structure of the browsing graph is more and more complete each day. At the 15th day, the browsing graph contains 6.12 million edges and 2.56 million vertices. From then on, the number of newly visited edges and vertices is relatively stable. Approximately 0.3 million new edges and 0.1 million new vertices appear on each subsequent day. Therefore, it takes approximately 15 days to construct a stable user browsing graph and subsequently, approximately 20% of newly visited Web sites are not included in *BG(V,E)* each day.

Step 2. According to Step 1, approximately 20% of newly visited sites would be missing if we adopt *BG(V,E)* for quality estimation (supposing *BG(V,E)* is daily updated). To determine whether this missing subset of newly visited sites affects quality estimation, we examined whether Web sites that are not included in the graph are indexed by search engines. If they are not indexed by search engines, it is not necessary to calculate their quality estimation scores because search engines will not require these scores. We sampled 30,605 pages from the sites that are visited by users but not included in *BG(V,E)* (approximately 1% of all visited pages in these sites) and checked whether they are indexed by four widely used Chinese search engines (Baidu.com, Google.cn, Sogou.com, Yahoo.cn). The experimental results are shown in Table 6 (SE1-SE4 is used instead of search engine names).

Table 6. Percentage of newly visited pages indexed by search engines

Search Engine	Percentage of pages indexed
SE1	8.65%
SE2	11.52%
SE3	10.47%
SE4	14.41%
Average	11.26%

The experimental results in Table 6 show that most of these pages (88.74% on average) are not indexed by search engines. It is not necessary for *BG(V,E)* to include these pages because search engine do not require their quality estimation scores.

Step 3. According to results of Step 1 and 2, we can calculate that only 2.2% (11.26% × 20%) of newly visited pages are both not included in *BG(V,E)* and required for quality estimation. Among the pages that are both indexed by search engines and visited by users, most will be included by *BG(V,E)* if this graph can be updated daily with new log data on browsing behavior. Therefore, it is appropriate to use *BG(V,E)* in quality estimation. Because *user-HG(V,E)* and *user-CG(V,E)* share the same vertex set with *BG(V,E)*, these constructed graphs are also not substantially affected by the problem of new visits to missing pages. Thus, these graphs are also appropriate for quality estimation.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

6.1 Experimental Setup

In Section 1, we assume that the user-accessed part of Web is more reliable than the parts that are never visited by users. On the basis of this assumption, we construct three different hyperlink graphs based on browsing behavior. To determine whether the constructed graphs outperform original Web graph in estimating page quality, we adopted two evaluation methods.

The first method is based on the ROC/AUC metric, which is a traditional measure in quality estimation research, such as "Web Spam Challenge"[6]. To construct a ROC/AUC test set, we sampled 2,279 Web sites randomly according to their frequencies of user visits and had two assessors annotate their quality scores. Approximately 39% of these sites were annotated as "high quality", 19% were "spam", and the others are "ordinary". After performing link analysis algorithms, each site in the test set was assigned a quality estimation score. We can evaluate the performance of a link analysis algorithm on the basis of whether it assigns higher scores to good pages and lower scores to bad ones.

[6] http://webspam.lip6.fr/

The second method is a pairwise orderedness test. This test was first proposed by Gyöngyi et al. [11] and is based on the assumption that good pages should be ranked higher than bad pages by an ideal algorithm. We constructed a pairwise orderedness test set composed of 782 pairs of Web sites. These pairs were annotated by product managers of a Web user survey company. It is believed that the pairwise orderedness show the two sites' differences in reputation. For example, both http://video.sina.com.cn/ and http://v.blog.sohu.com/ are famous video-sharing Web sites in China. However, the former site is more popular and receives more user visits, so the pairwise quality order is http://video.sina.com.cn/ > http:// v.blog.sohu.com/. If an algorithm assigns a higher score to http://video.sina.com.cn/, it passes this pairwise orderedness test. We use the accuracy rate to evaluate the performance of the pairwise orderedness test, which is defined as the percentage of correctly ranked Web site pairs.

With these two evaluation methods, we tested whether traditional hyperlink analysis algorithms perform better on *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)* than on the original Web graph. In addition, we also investigated whether a specifically designed link analysis algorithm for browsing graphs (such as BrowseRank) performs better traditional methods (such as PageRank and TrustRank).

First, we compared the performance of the link analysis algorithms on the four graphs (*BG(V,E)*, *user-HG(V,E)*, *user-CG(V,E)* and *whole-HG(V,E)*). Second, we compared the performance of PageRank, TrustRank, DiffusionRank and BrowseRank on *BG(V,E)*. The latter comparisons were only performed on *BG(V,E)* because BrowseRank requires users' stay time information, which is only applicable for *BG(V,E)*. In addition, to examine how the proposed userPageRank and userTrustRank algorithms perform, we compared their performances to that of the original

algorithms on both a user browsing graph and a social graph constructed with data from China's largest micro-blogging service provider weibo.com.

For TrustRank and DiffusionRank, a high quality page "seed" set must be constructed. In these experiments, we follow the construction method proposed by Gyöngyi *et al* in [11] and which is based on an inverse PageRank algorithm and human annotation. The inverse PageRank algorithm was performed on the whole Web graph, and we annotated the top 2000 Web sites ranked by inverse PageRank. Finally, 1153 high quality and popular Web sites were selected to compose the seed set. The parameters in our implementation of PageRank, TrustRank and Diffusion Rank algorithms are all tuned according to their original implementations [11, 25, 28]. The α parameters of PageRank and TrustRank algorithms are set to 0.85 according to [25] and [11]; and the iteration time are both set to 30 because that is enough for the results to converge. Parameters for the DiffusionRank algorithm are set as: $\gamma = 1.0$, $\alpha=0.85$, $M=100$ according to [28].

6.2 Quality Estimation with Different Graphs

With the four different hyperlink graphs shown in Table 5, we applied the PageRank algorithm and evaluated the performance of page quality estimation. The experimental results of high quality page identification, spam page identification and the pairwise orderedness test are shown in Figure 5. The performances of high quality and spam page identification are measured by the AUC value, whereas the pairwise orderedness test used accuracy as the evaluation metric.

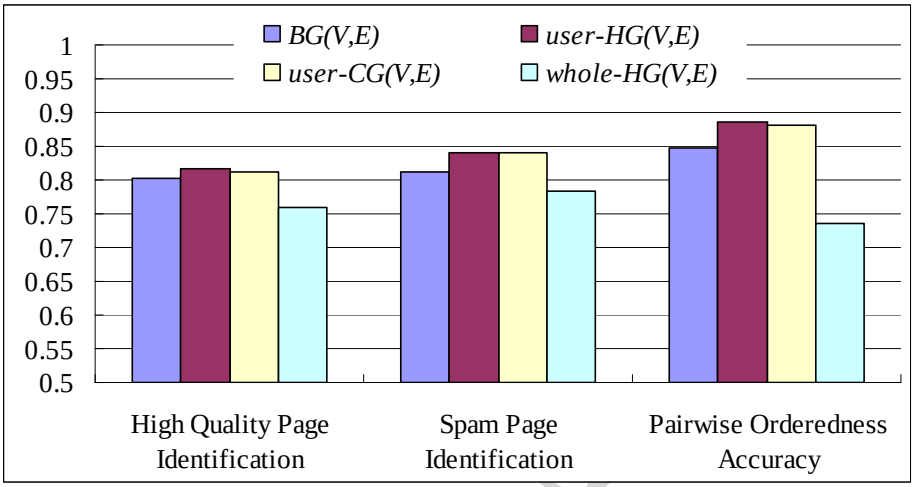

Figure 5. Quality estimation results with PageRank performed on *BG(V,E)*, *user-HG(V,E)*, *user-CG(V,E)* and *whole-HG(V,E)*

Figure 5 shows that that PageRank applied to the original Web graph (*whole-HG(V,E)*) performs the worst in all three quality estimation tasks. This result indicates that the graphs constructed by Algorithms 1-3 can more effectively estimate Web page quality than can the original Web graph. The improvements in performance associated with each of these three graphs are shown in Table 7.

Table 7. Performance improvements of the graphs constructed with Algorithms 1-3 compared to the original Web graph

Test Method	**Improvement compared with *whole-HG(V,E)***		
	BG (V,E)	***user-HG(V,E)***	***user-CG(V,E)***
High quality page identification	+5.69%	+7.55%	+7.12%
Spam page identification	+3.77%	+7.44%	+7.46%
Pairwise orderedness test	+15.14%	+20.34%	+19.67%

According to Table 7, the graphs constructed with information on browsing behavior outperform the original Web graph by approximately 5-25%. The adoption of user browsing behavior helps reduce possible noise in the original graph and makes the graph more reliable. This finding agrees with the results in [23] that *BG(V,E)* outperforms the original Web graph. It also validates our assumption proposed in Section 1 that the user-accessed part of Web is more reliable than the parts that are never visited by users.

According to Figure 5 and Table 7, among the three graphs constructed with user behavior information, *BG(V,E)* performs the worst, whereas *user-HG(V,E)* and *user-CG(V,E)* obtain very similar results. As described in Section 3.5, *BG(V,E)* contains fewer edges than the other two graphs. The retained links are on average more informative than the edges in the other graphs; however, this huge loss of edge data also compromises the page quality estimation. *User-HG(V,E)* and *user-CG(V,E)* share the same vertex set, and their edge sets are also very similar (only 7.97 million edges are added to *user-CG(V,E)*, making up 5.14% edges of the whole graph). Therefore, these two graphs perform similarly in page quality evaluation.

BG(V,E), *user-HG(V,E)* and *user-CG(V,E)* share the same vertex set, which is composed of all user-accessed sites recorded in Web-access logs. Although *BG(V,E)* contains the fewest edges of the four graphs, it still outperforms *whole-HG(V,E)*. This result shows that the selection of the vertex set is more important than the selection of the edge set. Reducing the unvisited nodes in the original Web graph can be an effective method for constructing hyperlink graph.

In Section 1, we show in Table 1 a list of Web sites which are ranked top according to PageRank scores on the original Web graph. We also find that some government Web sites (e.g. www.miiberan.gov.cn, www.hd315.gov.cn) are ranked quite high but fail to draw much user attention. These Web sites are authoritative and important but they should not be ranked so high because other similar government agency Web sites are generally ranked much lower. However, when we look into the results of PageRank performed on *BG (V,E)*, we find that the rankings of www.miiberan.gov.cn and www.hd315.gov.cn are more reasonable.

Table 8. PageRank ranking comparison of some government agency Websites on *whole-HG(V,E)* and *BG(V,E)*

	PageRank Ranking on *whole-HG(V,E)*	**PageRank Ranking on *BG(V,E)***

www.miibeian.gov.cn	5	23
www.hd315.gov.cn	2	117

According to Table 8, both www.miiberan.gov.cn and www.hd315.gov.cn are ranked lower according to PageRank on *BG(V,E)* than that on *whole-HG(V,E)*. They are also important resources according to algorithm on the user browsing graph but not as important as the top-ranked ones. We believed that the rankings on *BG(V,E)* give a better estimation of their quality according to both popularity and authority.

6.3 Quality Estimation with Different Link Analysis Algorithms

In [23], Liu et al. have shown that a specifically designed link analysis algorithm (BrowseRank) outperforms TrustRank and PageRank for both spam fighting and high quality page identification when the latter two algorithms are applied to the original Web graph. They explained that BrowseRank improves performance because it can better represent users' preferences than PageRank and TrustRank. However, it is still unclear whether this improvement comes from algorithm and model design or from the adoption of data on user behavior. Thus, we tested the performance of PageRank, TrustRank and BrowseRank on the same *BG(V,E)* graph. This comparison was only performed on *BG(V,E)* because the calculation of BrowseRank requires users' stay time information, which is applicable to *BG(V,E)* only.

PageRank performs better on *BG(V,E)* than on the original Web graph (Figure 5). Therefore, it is possible that the BrowseRank algorithm improves performance simply because it is performed on a graph constructed from data on user browsing behavior. The experimental results shown in Figure 6 validate this assumption. TrustRank performs the best in both spam page identification and high quality page identification, whereas PageRank performs slightly better than the other three algorithms in the pairwise orderedness test. The good performance of TrustRank might come from the prior information stored in the "seed" set.

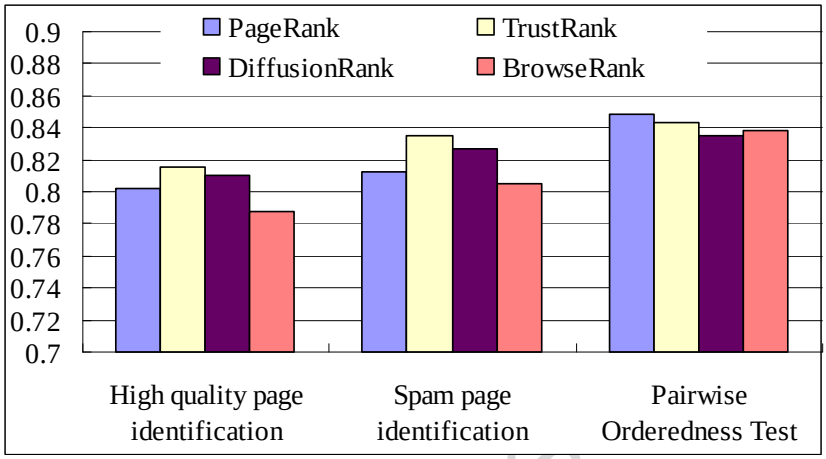

Figure 6. Results of quality estimation with different link analysis algorithms on *BG(V,E)*

According to the results, TrustRank outperforms BrowseRank by 4.12% and 2.84% in high quality and spam page identification tasks, respectively. The performance improvements are small but demonstrate that the TrustRank algorithm can also be very effective on *BG(V,E)*. The PageRank algorithm also performs no worse than BrowseRank on any of these tests. This result means that the performance improvement by the BrowseRank algorithm reported in [23] comes both from algorithm design and, perhaps more importantly, from the adoption of information on user browsing behavior. Additionally, PageRank and TrustRank are more efficient than BrowseRank because they do not require collecting information on users' stay time.

These results and examples demonstrate that although BrowseRank is specially designed for *BG(V,E)*, it does not perform better than PageRank, TrustRank or DiffusionRank applied to *BG(V,E)*. BrowseRank favors the pages where users stay longer, but stay time does not necessarily indicate quality or user preference. Compared with the algorithm design, the incorporation of information on user browsing behavior in the construction of link graphs is perhaps more important.

6.4 UserPageRank and UserTrustRank on User Browsing Graph

In Section 3, we proposed the userPageRank and userTrustRank algorithms, which modify the

original algorithms by estimation of $P(X_i \Rightarrow X_j)$ according to user browsing information recorded in *BG(V,E)*. To examine the effectiveness of these algorithms, we compared their performance with the original PageRank/TrustRank algorithms (Figure 7).

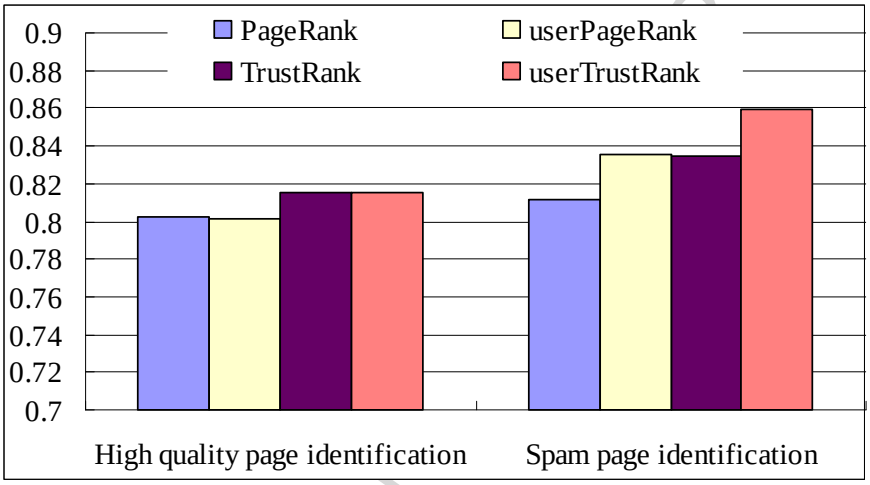

Figure 7. Quality estimation results with the original PageRank/TrustRank and userPageRank/userTrustRank algorithms on *BG(V,E)*

The modified algorithms perform slightly better than the original algorithms. They perform almost equivalently in high quality page identification and perform slightly different in spam page identification. For both PageRank and TrustRank algorithms, the modified algorithms outperform the original ones by approximately 3% in spam identification. Examining several test cases, we find that this performance improvement comes from modification to the algorithms.

An example is the spam site whose URL is http://11sss11xp.org/. Among the 2279 Web sites in the ROC/AUC test set, it is ranked 1030th by the original TrustRank algorithm and 1672nd by the userTrustRank algorithm. Because a spam site should be assigned a low ranking position, userTrustRank performs better for this test case. We investigated the hyperlink structure of this site to analyze why the modified algorithm performs better.

Table 9. Web sites that link to a spam site (http:// 11sss11xp.org/) in *BG(V,E)*

Source Web site	Destination Web site	#User Visits
http://web.gougou.com/	http://11sss11xp.org/	3
http://image.baidu.com/	http://11sss11xp.org/	1
http://www.yahoo.cn/	http://11sss11xp.org/	1
http://domainhelp.search.com/	http://11sss11xp.org/	1
http://my.51.com/	http://11sss11xp.org/	1

Table 10. Information on sites that connect to a spam site (http://11sss11xp.org/)

Site	#Out-link	#User Visits
www.yahoo.cn	35,000	208,658
my.51.com	86,295	19,443,717
image.baidu.com	148,611	8,218,706

In Tables 9 and 10, we can see that this site receives many in-links from search engines (such as www.yahoo.cn and image.baidu.com). This phenomenon can be explained because spam sites are designed to achieve unjustifiably favorable rankings in search engines. This spam site also receives in-links from several Web 2.0 sites, such as my.51.com, which is a blog service site. With the original TrustRank algorithm, trust scores of the original sites should be evenly divided between their outgoing links. In contrast, for userTrustRank, trust scores are assigned by estimating $P(X_i \Rightarrow X_j)$, the probability of visiting site X_j after visiting X_i. Because this site is a spam site that users generally do not visit, $P(X_i \Rightarrow X_j)$ for this site should be low. For example, the site www.yahoo.cn has 35,000 outgoing links in *BG(V,E)*. Altogether, 208,658 user clicks are performed on these outgoing links, and only one of them links to 11sss11xp.org. With the original TrustRank algorithm, the spam site receives 1/35000 of Yahoo's trust score, whereas userTrustRank only assigns 1/208658 of the corresponding score to this spam site. We can see that userTrustRank divide a page's trust score according to counts of users' visits, and this adaptation can help identify spam sites.

6.5 UserPageRank and UserTrustRank on Social Graph

In order to further examine the performance of userPageRank and userTrustRank algorithms, we also constructed a social graph as described in Section 3 and see how they performs on it.

The data was collected in September, 2011 from weibo.com, which is China's largest social network service provider. Information of 2,631,342 users and about 3.6 billion relationships were collected. To the best of our knowledge, it is one of the largest corpuses in social network studies. Information recorded in our data set is shown in Table 11.

Table 11. Information recorded in the collected micro-blogging data

Information	Explanations
User ID	The unique identifier for each user
User name	The name of the user
Verified sign	Whether the user's identification is verified by weibo.com
Followees	The ID list that are followed by the user
Followers	The ID list that follow the user
Tags	A list of keywords describing the user's interests with the purpose of self-introduction

As described in Section 3, the userPageRank and userTrustRank requires the estimation of $P(X_i \Rightarrow X_j)$ as prior knowledge. In social graph, we adopted the number of common tags as a sign of closeness between users. We believe that the assumption is reasonable because the following relationships between users with many common interests should be more reliable than those not. Therefore, the weight of an edge in the social graph equals to the number of common tags between nodes it connects. After performing userPageRank and userTrustRank algorithms on the weighted social graph, social influence estimation results are shown in Figure 8.

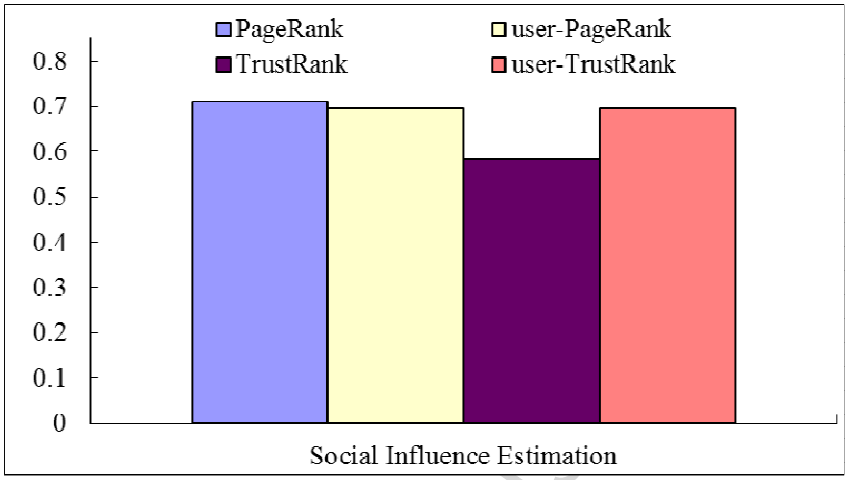

Figure 8. Social influence estimation results with the original PageRank/TrustRank and userPageRank/userTrustRank algorithms on social graph of weibo.com

Figure 8 shows the AUC performances of different influence estimation algorithms on the social graph. We use the users with "Verified sign" as more influent ones in our evaluation because their identity has been verified by weibo.com and according to the verification policy[7], only "authoritative" person or organizations will be verified. For the seed set of TrustRank and userTrustRank, we select 100 people from "Weibo hall of fame[8]" which is composed of famous people in certain fields such as entertainment, politics, techniques and so on.

According to results shown in Figure 8, we see that the performance of PageRank, userPageRank and userTrustRank are similar to each other while TrustRank performs the worst among all algorithms. Although the AUC performance of PageRank is almost the same as userPageRank and userTrustRank, we find that these algorithms give quite different rankings. The top results of the algorithms in Table 12 show that both PageRank and TrustRank put famous entertainment stars (such as Xidi Xu, Chen Yao and Mi Yang) at the top of their result lists. Meanwhile, userPageRank and userTrustRank favor accounts which post interesting jokes

[7] http://weibo.com/verify

[8] http://weibo.com/pub/star

or quotations (such as joke selection and classic quotations).

Table 12. Top results of PageRank, TrustRank, userPageRank and userTrustRank algorithms on the social graph of weibo.com

Rank	PageRank	userPageRank	TrustRank	userTrustRank
1	Kangyong Cai	Joke Selection	Kangyong Cai	Kangyong Cai
2	Xidi Xu	Kangyong Cai	Mi Yang	Joke Selection
3	Cold Joke Selection	Classic Quotations	Na Xie	Xiaoxian Zhang
4	Chen Yao	Cold Joke Selection	Weiqi Fan	Classic Quotations
5	Xiaogang Feng	Global Fashion	Lihong Wang	Cold Joke Selection

The differences in top ranked results are caused by the fact that although the entertainment stars have many followers, a large part of these followers do not share same tags with the stars. This is because many of the stars do not list any tags on their accounts such as Xidi Xu and Chen Yao. People follow the accounts such as joke selection and classic quotations because they actually provide interesting information and influent people. Therefore, we believe that userPageRank and userTrustRank algorithms give more reasonable estimation of social influence.

7. CONCLUSION AND FUTURE WORKS

Page quality estimation is one of the greatest challenges for search engines. Link analysis algorithms have made progress in this field but encounter increasing challenges in the real Web environment. In this paper, we analyze user browsing behavior and proposed two hyperlink analysis algorithms based on “surfing with prior knowledge” model instead of the random surfer model. We also construct reliable link graphs in which this browsing behavior information is embedded. Three construction algorithms are adopted to construct three different kinds of link graphs, *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)*. We examined the structure of these graphs and found that they inherit characteristics, such as power law distributions of in-degrees and

out-degrees, from the original Web graph. The evolution of these graphs is also studied, and they are found to be appropriate for page quality estimation by search engines.

The experimental results show that the graphs constructed with browsing behavior data are more effective than the original Web graph in estimating Web page quality. PageRank on *BG(V,E)*, *user-HG(V,E)* and *user-CG(V,E)* outperforms PageRank on the whole Web graph. In addition, *user-HG(V,E)* and *user-CG(V,E)* work better than *BG(V,E)*, probably because the construction process of *BG(V,E)* omits too many meaningful hyperlinks. We also found that PageRank, TrustRank and DiffusionRank perform as well as (or even better than) BrowseRank when they are performed on the same graph (*BG(V,E)*). This result reveals that the incorporation of user browsing information is perhaps more important than the selection of link analysis algorithms. Additionally, the construction of user browsing graphs introduces more information. Thus, it is possible to modify the original TrustRank/PageRank algorithms by estimating the importance of outgoing links. The modified algorithms (called userPageRank and userTrustRank) show better performance in both Web spam identification and social influence estimation.

Although the Web / micro-blogging collections and data on user browsing behavior are collected on Chinese Web environment, the algorithms are not specially designed for the specific collection. Therefore, they should not behave significantly differently in a multi-language collection as long as reliable data sources can be provided.

Several technical issues remain, which we address here as future work:

First, Web pages that are visited by users only comprise a small fraction of pages on the Web. Although it has been found that most pages that users need can be included in the vertex set of

BG(V,E), search engines still need to keep many more pages in their index to meet all possible user needs. To estimate quality of these pages, we are planning to predict user preferences for a certain page by using the pages that users previously visited as training set. If we can calculate the probability that a Web page will be visited by users in the future, this information will help construct a large-scale, credible link graph not limited by data on user behavior.

Second, the evolution of the user browsing graph can be regarded as a combination of the evolution of both the Web and Web users' interests. In this paper, we analyzed the short term evolution (a period of 60 days) of the graph. We are considering collecting long-term data to determine how the evolutionary process reflects changes in users' behavior and interests.

ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A207) of China. In the early stages of this work, we benefited enormously from discussions with Yijiang Jin. We thank Jianli Ni for kindly offering help in data collection and corpus construction. We also thank Tao Hong, Fei Ma, Shouke Qin from Baidu.com and the anonymous referees of this paper for their valuable comments and suggestions.

Biographical Note

Yiqun Liu, Male, borned in January, 1981. I recieved my bachelor and Ph.D. degrees from Dept. of Computer science and technology of Tsinghua University in July, 2003 and July, 2007, respectively. I am now working as a assistant professor in Tsinghua University and undergraduate mentor for the C.S.&T. Department. My research interests includes Web information retrieval, Web user behavior analysis and performance evaluation of on-line services. Most of my recent works and publications can be found at my homepage(http://www.thuir.cn/group/~YQLiu/).

REFERENCES

[1] B. Amento, L. Terveen, W. Hill, Does authority mean quality? Predicting expert quality ratings of Web documents. In Proc. of 23rd ACM SIGIR Conference (2000) 296-303

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates, Using Rank Propagation and Probabilistic Counting for Link Based Spam Detection. Proceedings of the Workshop on Web Mining and Web Usage Analysis (2006).

[3] M.Bendersky, W. Bruce Croft, Y. Diao, 2011. Quality-biased ranking of web documents. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 95-104.

[4] M. Bilenko, R. W. White, Mining the search trails of surfing crowds: identifying relevant Web sites from user activity. In Proc. the 17th WWW Conference. (2008) 51-60.

[5] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN System 30 (1998), 107-117.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the Web. Computer Networks 33 (2000) 309–320.

[7] M. Chau, H. Chen, A machine learning approach to web page filtering using content and structure analysis, Decision Support Systems, 44(2) (2008) 482-494.

[8] N. Craswell, D. Hawking, S. Robertson, Effective site finding using link anchor information. Proceedings of the 24th ACM SIGIR Conference (2001) 250-257.

[9] D. Donato, L. Laura, S. Leonardi, S. Millozzi, The Web as a graph: How far we are. ACM Transaction on Internet Technology 7(1) (2007), 4.

[10] X. Fang, C. W. Holsapple, An empirical study of web site navigation structures' impacts on

web site usability, Decision Support Systems, 43(2) (2007) 476-491.

[11] Z. Gyöngyi, H. Garcia-Molina, J. Pedersen, Combating web spam with trustrank. In Proceedings of the Thirtieth international VLDB Conference (2004) 576-587.

[12] T. Haveliwala, Efficient computation of PageRank. Technical Report, Stanford University, 1999. http://dbpubs.stanford.edu/pub/1999-31.

[13] T. Haveliwala, 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search. IEEE Transaction on Knowledge and Data Engineering. 15, 4, 784-796.

[14] T. Haveliwala, S. Kamvar, G. Jeh, An analytical comparison of approaches to personalizing PageRank. Stanford Technical Report, http://ilpubs.stanford.edu:8090/596/

[15] M. R. Henzinger, R. Motwani, C. Silverstein, 2002. Challenges in web search engines. SIGIR Forum 36, 2 (Sep. 2002), 11-22.

[16] A. Jacob, C. Olivier C. Carlos, WITCH: A New Approach to Web Spam Detection. Yahoo! Research Report No. YR-2008-001. (2008).

[17] Y. Kang, Y. Kim, Do visitors' interest level and perceived quantity of web page content matter in shaping the attitude toward a web site? Decision Support Systems, 42(2) (2006) 1187-1202.

[18] R. Kaul, Y. Yun, S. Kim, Ranking billions of web pages using diodes. Communications of the ACM, 52(8) (2009), 132-136.

[19] J. M. Kleinberg, Authoritative sources in a hyperlinked environment. Journal of ACM 46(5) (1999) 604-632.

[20] V. Krishnan, R. Raj, Web Spam Detection with Anti-TrustRank. In the 2nd International Workshop on Adversarial Information Retrieval on the Web, (2006) 3.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Core algorithms in the CLEVER system, in ACM Transactions on Internet Technology 6(2) (2006) 131-152.

[22] Y. Liu, F. Chen, W. Kong, H. Yu, M. Zhang, S. Ma, L. Ru. Identifying Web Spam with the Wisdom of the Crowds. ACM Transaction on the Web. Volume 6, Issue 1, Article No. 2, 30 pages. March 2012.

[23] Y. Liu, B. Gao, T. Liu, Y. Zhang, Z. Ma, S. He, H. Li, BrowseRank: letting web users vote for page importance. In Proc. of 31st ACM SIGIR Conference (2008) 451-458.

[24] Y. Liu, M. Zhang, R. Cen, L. Ru, L., S. Ma, Data cleansing for Web information retrieval using query independent features. Journal of the American Society for Information Science and Technology, 58(12) (2007), 1884-1898.

[25] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the Web. Stanford Technical Report. (1999) http://ilpubs.stanford.edu:8090/422/.

[26] H. Wu, M. Gordon, K. DeMaagd, W. Fan, Mining web navigations for intelligence, Decision Support Systems, 41(3) (2006) 574-591.

[27] D. Xu, Y. Liu, M. Zhang, L. Ru, S. Ma. Predicting Epidemic Tendency through Search Behavior Analysis. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11) (Barcelona, Spain). pp. 2361-2366.

[28] H. Yang, I. King, M.R. Lyu, DiffusionRank: A Possible Penicillin for Web Spamming. In Proc. of 30th ACM SIGIR Conference (2007) 431-438.

[29] B. Zhou, Y. Liu, M. Zhang, Y. Jin, S. Ma, Incorporating Web Browsing Information into Anchor Texts for Web Search, Information Retrieval Volume 14, Issue 3: 290-314, 2011.