User Browsing Graph: Structure, Evolution and Application[1]

Yiqun Liu, Min Zhang, Shaoping Ma, Liyun Ru
State Key Lab of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China P.R.
yiqunliu@tsinghua.edu.cn

ABSTRACT
This paper focuses on 'user browsing graph' which is constructed with users' click-through behavior modeled with Web access logs. User browsing graph has recently been adopted to improve Web search performance and the initial study shows it is more reliable than hyperlink graph for inferring page importance. However, structure and evolution of the user browsing graph haven't been fully studied and many questions remain to be answered. In this paper, we look into the structure of the user browsing graph and its evolution over time. We try to give a quantitative analysis on the difference in graph structure between hyperlink graph and user browsing graph, and then find out why link analysis algorithms perform better on the browsing graph. We also propose a method for combining user behavior information into hyper link graph. Experimental results show that user browsing graph and hyperlink graph share few links in common and a combination of these two graphs can gain good performance in quality estimation of pages.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search.

General Terms
Experimentation, Measurement.

Keywords
Link structure analysis, Brows graph， User behavior, Search Engines.

1. INTRODUCTION
Web page quality estimation is considered as one of the major challenges for Web search engines [1]. For contemporary search engines, estimating page quality plays an important role in crawling, indexing and ranking processes.

Currently, the task of page quality estimation is usually based on hyperlink structure analysis of the Web. The success of PageRank[2] and other hyperlink analysis algorithms such as HITS[3] prove that it is possible to evaluate the importance of a Web page query-independently. In these analyses, two basic assumptions are concluded by [4]: *recommendation assumption* and *topic locality assumption*. It is assumed that if two pages are connected by a hyperlink, the page linked is recommended by the page which links to it (recommendation) and the two pages share a similar topic (locality). Hyperlink analysis algorithms adopted by both commercial search engines (such as [5]) and researchers (such as [6]) all rely on these two assumptions. However, contemporary WWW is filled with spam links and advertising links so the assumptions as well as the consequent algorithms don't work very well in this new situation.

Recently, the *wisdom of the crowd* is paid much attention in Web search researches, e.g. [7], [8] and [9]. In their work, users' browsing behavior is usually considered as implicit feedback information for page relevance and importance. For example, Liu et. al. constructed 'user browsing graph' with search log data [10]. They proposed a page importance estimation algorithm called BrowseRank which performs on the user browsing graph. It is believed that the link structure in user browsing graph is more reliable than hyperlink graph because users actually follow links in the browsing graph.

Liu's initial study shows that the BrowseRank algorithm works better than hyperlink analysis algorithms such as PageRank and TrustRank [10]. However, there remain many questions to be answered for the user browsing graph. We can infer that user browsing graph is different from the hyperlink graph in some aspects, but how the structures of these two graphs differ from each other? How the user browsing graph evolves over time and even every day? We've known that BrowseRank performs better than PageRank and TrustRank algorithms when the latter two algorithms are performed on hyperlink graph. Then on the other hand, how these algorithms perform on the user browsing graph?

In this paper, we look into the structure and the evolution of the user browsing graph. And furthermore we study how traditional link analysis algorithms perform on the user browsing graph. We try to answer the following questions:

1. How does the user browsing graph differ from the hyperlink graph? How many edges are overlapped, reduced and complemented, respectively?
2. How does the user browsing graph evolve over time? How much data are necessary to build a stable graph structure and how the graph changes day by day?
3. How traditional hyperlink analysis algorithms perform on the user browsing graph? Do they perform better on the browsing graph than on the hyperlink graph?

The rest of the paper is organized as follows: Section 2 introduces user behavior dataset. Section 3 describes the structure and evolution of user browsing graph. Experimental results of applying link analysis algorithms on user browsing graph are reported in Section 4. Conclusions and future work are given in

[1] This work was supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141).

Section 5.

2. USER BEHAVIOR DATA SET

With the development of search engine, Web browser toolbars become more and more popular recently. Lots of search engines develop toolbar software to attract more user visits, e.g. Google, Yahoo and Live search. Web users usually adopt toolbars to get instant access to search engine services and to get browser enhancement such as pop-up window blocking and download acceleration. In order to provide value-add services to users, most toolbar services also collect anonymous click-through information from users' browsing behavior. Previous work such as [8] adopts this kind of click-through information to improve ranking performance. Our previous work [10] 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 search toolbar because this kind of data sources collect user behavior information at low cost without interrupting users' browsing behavior. Information recorded in these logs is shown in Table 1.

Table 1. Information recorded in Web access logs

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

From Table 1 we can see that no privacy information was included in the log data. The information shown can be easily recorded using browser toolbars by commercial search engine systems. Therefore it is practical and feasible to obtain these types of information and to apply them in the construction of user browsing graph. With the help of a widely-used commercial Chinese search engine, Web access logs were collected from Aug.3rd, 2008 to Oct 6th, 2008. Over 2.8 billion click-through events on 8.5 million Web sites were recorded in these logs.

3. STUCTURE AND EVOLUTION OF THE USER BROWSING GRAPH

3.1 The Construction Process

User browsing graph is constructed with users' behavior data recorded in Web access logs. We can use *UG(V,E)* to represent the user browsing graph in which *V* is the vertex set and *E* is the edge set. The construction process can be described as follows:

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 $(A,B) \notin E$
   $E = E \cup \{(A,B)\}, Weight(A,B) = 1;$
   else
   $Weight(A,B)++;$

After the construction process, *V* includes all Web pages visited by users during the time period in which access logs were collected; and *E* records the users' browsing behaviors. Each edge in *E* is also assigned a weight value which represents how many times Web users visit B from A.

3.2 The Structure of User Browsing Graph

For each edge *(A,B)* from *UG(V,E)*, there exists some user who visits Web page *B* by following links located on page *A*. It means that *UG(V,E)* can be regarded as a subset of the hyperlink graph on the Web (named *HG(V,E)*), because the latter graph includes all hyperlinks and pages on the Web. However, *HG(V,E)* is constructed with the information collected by Web crawlers and it is not possible for any crawler to collect hyperlink graph of the whole Web because it is too huge and changing so fast. It is also not necessary to retain the whole graph structure because Web users can only view part of it due to limited time. Therefore, it is likely that the *HG(V,E)* graph maintained by search engines may not contain all vertexes and edges in *UG(V,E)*.

In order to find out the differences between the two graphs, we constructed site-level *UG(V,E)* with web access data from Aug.3rd, 2008 to Sept.2nd, 2008 (totally 30 days). The vertex set contains 4,252,495 Web sites and the edge set contains 10,564,205 edges. We constructed site-level instead of page-level *UG(V,E)* because we believe that the site-level graph is more stable and we also want to avoid the data sparsity problem.

With the help of the same search engine which collected Web access log for us, we obtained hyperlink graph constructed by them which contains hyperlink relations of over 3 billion Web pages. Then we extracted a sub-graph for all the Web sites in *V* of *UG(V,E)*. The sub-graph is the hyperlink graph for these Web sites and we found that although the hyperlink graph contains almost the same Web sites as the user browsing graph, structures of these two graphs are significantly different. Table 2 shows differences in the edge set.

Table 2. Differences between the hyperlink graph and the user browsing graph in the edge set

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

From Table 2 we can see that user browsing graph and hyperlink graph share only a small proportion of edges in common. There are only less than 1/4 of pages in user browsing graph that also appear in hyperlink graph. The percentage of common pages in *HG(V,E)* is only 1.86% because it has much more pages.

We can see from Table 2 that most (98.14%) of the links in the hyperlink graph are not actually clicked by users. This can be explained by the fact that Web pages usually provide too many hyperlinks for the user to click. We can also find that over 75% of the links in the user browsing graph are not included in the hyperlink graph. When we look into these links, we found that most of these links come from user clicks on search engines. Table 3 shows the number of search engine oriented edges in *UG(V,E)* that are not included in *HG(V,E)*.

Table 3. Number of search engine oriented edges that are not included in *HG(V,E)*

Search Engine	Edges that are not included in *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	*3,341,749 (41.92% of all edges in UG(V,E))*

We can see from Table 2 and 3 that among the links only appearing in *UG(V,E)* (totally 7.97 million edges), over 3.34 million are oriented by the five most frequently-used Chinese search engines. This number covers 41.92% of all edges in *UG(V,E)*. Web users click lots of links on search engine result pages, but few of these links are collected by crawlers. We believe that these links should be recorded because they link to valuable pages that are both recommended by search engines and selected by users. It is not possible for Web crawlers to collect all links from search result pages because each search request results in such a page and the number would be quite huge.

Another important type of links that appear only in *UG(V,E)* are hyperlinks that are clicked in users' password-protected sessions. For example, login authorization is sometimes needed to visit one's blog Web pages. After login process it is possible for Web users to 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 contents of these protected Web pages.

From the above statements, we can see that user browsing graph is different from hyperlink graph in at least two ways: Compared to hyperlink graph, a large part of edges (98.14% of *E* in *HG(V,E)*) are reduced in the browsing graph because they are not clicked by any user. At the same time, some links are added which are difficult or impossible to be collected by Web crawlers. This makes these two graphs significantly different from each other.

3.3 The Evolution of User Browsing Graph

In [10], user browsing graph is used to calculate the importance of Web pages. For practical applications, an important issue is whether the page importance scores calculated off-line can be adopted for on-line search process. If pages needed by users are not included in the user browsing graph, it is impossible to calculate their importance scores and search engines may not be able to show these pages to users. Therefore, it is important to find out how user browsing graph evolves over time. The graph cannot cover all pages that are required by users because new pages appear from time to time. However, it is acceptable if only a small proportion of newly-appeared pages are not included in user browsing graph. Figure 1 shows evolution of the user browsing graph in the time period during which the dataset was prepared.

In Figure 1, we show how *V* and *E* in *UG(V,E)* evolves over time. We can see that on the first day all edges and vertexes were newly-appeared because both *V* and *E* are empty sets. From the second day to the 15th day (approximately), the percentage of newly-appeared edges and vertexes drops day by day. After the 15th day, about 25% of the edges and 30% of the vertexes appeared each day are new ones to the user browsing graph.

During the first 15 days, the percentage of newly-appeared edges and vertexes drops because the structure of the browsing graph is more and more complete day by day. At the fifteenth day, the browsing graph contains 6.12 million edges and 2.56 million vertexes. From then on, the amount of newly-appeared edges and vertexes are relatively stable and about 0.3 million new edges and 0.1 million new vertexes appear each day. Therefore, it takes about 15 days to construct a stable user browsing graph and after that, small part (less than 5%) of the graph changes each day. Because of the relatively small size of user browsing graph, daily reconstruction of the user browsing graph is possible. It means that most of the pages that users need can be included in the graph.

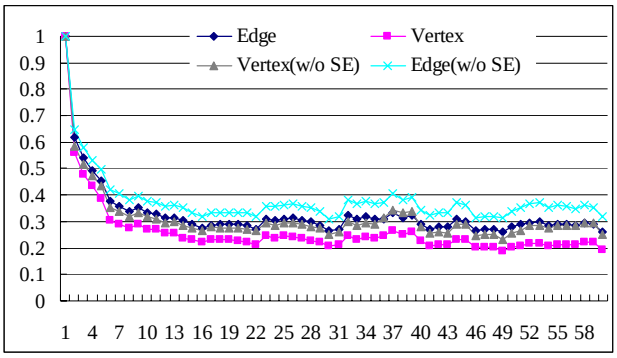

Figure 1. Evolution of the user browsing graph from Aug.3rd, 2008 to Oct 6th, 2008. Category axis: order of the day with Aug.3rd, 2008 as the 1st day. Value axis: percentage of newly appeared edges/vertexes on the corresponding day.

According to Section 3.2, links oriented by search engine result pages make up large part of the browsing graph; therefore, we look into this part of graph to see whether it changes faster or slower than the other parts. Figure 1 shows that without search engine oriented links, the percentages of newly-appeared edges/vertexes rise up to 30% and 35%, respectively. It means user clicks from search result pages don't change as fast as other user clicks. It is necessary to retain this kind of links in the browsing graph because they provide valuable information (See Section 3.2) and they compose a relatively stable part of the graph.

4. EXPERIMENTS AND RESULTS

From previous work in [10], we can see specifically-designed link analysis algorithm (named BrowseRank) can perform better on user browsing graph in page quality estimation than state-of-the-art link analysis algorithms which performs on hyperlink graphs. We want to find out whether this improvement comes from algorithm design or the differences in graph structure. We also want to know how hyperlink analysis algorithms such as PageRank [2] and TrustRank [11] perform on user browsing graph. If these algorithms perform better on user browsing graph, it is possible that browsing graph can replace hyperlink graph for page quality estimation because hyperlink graph is much huger in size and has brought much difficulty in computation complexity and information storage. We didn't compare the performance of BrowseRank with PageRank and TrustRank on user browsing graph because we are not able to obtain stay time information (which is necessary for BrowseRank calculation) from our Web access logs described in Section 2.

In order to answer these questions, we build four Web graphs and compares how PageRank and TrustRank algorithms perform on them. Details of these four graphs are shown in Table 4.

Table 4. User browsing graphs and hyperlink graphs constructed in our experiments

Graph	Description
User Browsing Graph *UG(V,E)*	Constructed with web access data from Aug.3rd, 2008 to Sept.2nd, 2008.
Hyperlink Graph *whole-HG(V,E)*	Constructed with over 3 billion pages (all pages in a certain search engine's index) and all hyperlinks among them

Hyperlink Graph *extracted-HG(V,E)*	Vertexes are from *UG(V,E)*. Edges among them are extracted from hyperlink relations in *whole-HG(V,E).*
Combined Graph *CG(V,E)*	Vertexes are from *UG(V,E)*. Edges among them are from *UG(V,E)* combined with those from *extracted-HG(V,E).*

After performing PageRank and TrustRank algorithms on these four graphs, we sampled 1680 Web sites according to user visit count and had 2 assessors annotate their quality scores. About 39% of these sites are annotated as "high quality"; 19% are "spam" and the others are "ordinary". After the annotation, we choose ROC curves and corresponding AUC values to evaluate the performance of link analysis algorithms. It is a useful technique for organizing classifiers and it is adopted by several quality estimation researches such as [12]. Table 5 and 6 show the performances of link analysis algorithms on different graphs.

Table 5. AUC/ROC values for high quality page identification performance of link analysis algorithms

	PageRank	TrustRank
UG(V,E)	0.84868	0.92032
whole-HG(V,E)	0.84113	0.85737
extracted-HG(V,E)	0.86960	0.91626
CG(V,E)	0.86756	0.91846

Table 6. AUC/ROC values for Web spam page identification performance of link analysis algorithms

	PageRank	TrustRank
UG(V,E)	0.87666	0.84627
whole-HG(V,E)	0.73659	0.73659
extracted-HG(V,E)	0.84686	0.84554
CG(V,E)	0.88014	0.88198

From Table 5 and 6 we can find several interesting results. The first is that among the four link graphs, link analysis algorithms performs the worst on *whole-HG(V,E)* (the whole hyperlink graph). This experimental result accords with the conclusion in [10] that link analysis algorithm performs better on user browsing graph than on the whole hyperlink graph. Another finding is that the sub-graph of *whole-HG(V,E)* (*extracted-HG(V,E)*) is better at identifying both high quality and spam pages than the whole graph. *Whole-HG(V,E)* is composed of much more hyperlinks than *extracted-HG(V,E)* but it results in performance loss in page quality estimation. This may be explained by the fact that *extracted-HG(V,E)* can be regarded as the user-accessed part of *Whole-HG(V,E)*. It shares the same vertex set of *UG(V,E)* and therefore reduces possible noises in the whole hyperlink graph.

Another result is that although *UG(V,E)* and *extracted-HG(V,E)* share only a small proportion of edges in common (see Table 2), they get similar performances in page importance estimation. A combination of *UG(V,E)* and *extracted-HG(V,E)* which contain all edges in them (*CG(V,E)*) result in best results. It means both user browsing graphs and hyperlink graph can provide useful information that the other graph doesn't contain.

From these results, we find that information recorded in user browsing graph is important for page quality estimation. Vertex set of *UG(V,E)*, *extracted-HG(V,E)*, *CG(V,E)* is the same and it is the user accessed part of Web. Focusing in this part of Web can reduce possible noises in hyperlink graph. As for the edge set, both hyperlink relations and user browsing relations are important and a combination can improve performance.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we look into structure and evolution of user browsing graph. First, quantitative analyses have been made to show the differences between the hyperlink graph and the user browsing graph in terms of structure and links. Second, the evolution of the graph has been studied, and experimental analysis shows that about 15 days data is enough for constructing a stable graph. Furthermore, the performances of traditional hyperlink analysis algorithms are studied in the user browsing graph. And finally, we proposed a combination algorithm to integrate hyperlink relation and user browsing relation to build a novel web page structure, which improves the performance of page quality estimation.

Future study will focus on improving search effectiveness and efficiency with link analysis algorithms performing on user browsing graphs.

6. REFERENCES

[1] Henzinger, M.R., Motwani, R. & Silverstein, C. (2003). Challenges in Web Search Engines (pp. 1573-1579). In the 18th International Joint Conference on Artificial Intelligence.

[2] Brin S. & Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane.

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

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

[5] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Stanford Digital Library Technologies Project.

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

[7] Fuxman, A., Tsaparas, P., Achan, K., and Agrawal, R. 2008. Using the wisdom of the crowds for keyword generation. In Proceeding of the 17th WWW Conference. 61-70.

[8] Bilenko, M. and White, R. W. 2008. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In Proceeding of the 17th WWW Conference. 51-60.

[9] Liu, Y., Cen, R., Zhang, M., Ma, S., and Ru, L. 2008. Identifying web spam with user behavior analysis. In the 4th international Workhop on Adversarial information Retrieval on the Web. AIRWeb '08. ACM, New York, NY, 9-16.

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

[11] Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. 2004. Combating web spam with trustrank. In Proceedings of the Thirtieth international VLDB Conference. Vol. 30. 576-587.

[12] Web Spam Challenge Website: http://webspam.lip6.fr/