

A Unified Framework for Web Link Analysis

Zheng Chen¹, Li Tao¹, Jidong Wang¹, Liu Wenyin², Wei-Ying Ma¹

¹Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R. China

{zhengc, i-jdwang, wyma}@microsoft.com

² Dept. of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
csluiwy@cityu.edu.hk

Abstract

Web link analysis has been proved to provide significant enhancement to the precision of web search in practice. Among existing approaches, Kleinberg's HITS and Google's PageRank are the two most representative algorithms that employ explicit hyperlinks structure among web pages to conduct link analysis, and DirectHit represents the other extreme that takes the user's access frequency as implicit link to the web page for counting its importance. In this paper, we propose a novel link analysis algorithm which puts both explicit and implicit link structures under a unified framework, and show that HITS and DirectHit are essentially the two extreme instances of our proposed method. One important advantage of our method is its ability to analyze not only the hyperlinks between web-pages but also the interactions between the users and the Web at the same time. The importance of web-pages and users can reinforce each other to improve the Web link analysis. Compared with traditional HITS and DirectHit algorithms, our method further improves the search precision by 11.8% and 25.3%.

1. Introduction

Finding useful information efficiently on the Web is becoming more and more difficult nowadays. In most cases, a user inputs some keywords to a search engine to find his desired information. Since "full text search" technology is widely adopted by most search engines currently, a large amount of documents containing the same keywords inputted by the user will be retrieved. It is a time consuming job for the user to go through the results to find out his really desired information. Many companies and researchers try to work out solutions to improve the precision of search engines. One of the representative solutions is to re-ranking the retrieved documents by their importance [1][8][12], which is calculated by analyzing the link between Web pages. Web link analysis [1][2][3][4][5][6][7][8][13] has been proved to reach higher precision than full text search in practice.

According to the type of web link, link analysis approaches can be classified into two categories, "explicit link analysis" and "implicit link analysis". The so-called

"explicit link" stands for the hyperlinks embedded in the web-page. It has been demonstrated that utilizing the hyperlink information can make great improvement to the performance of Web search. Kleinberg [1] analyzed hyperlinks among web-pages and summarized the concepts of authoritative page and hub page. Furthermore, he observed that these two kinds of web-pages generally reinforce each other and then presented the basic HITS algorithm [1]. Since the link structure of the web-pages is constructed by the website editors, the Kleinberg's algorithm calculates the importance of the web-pages from the editor's view. The assumption of the "explicit link analysis" is that users agree with the editor/author of the web pages in terms of the link structure. It works well in many cases but with a few exceptions. Another type of links is only implicitly imagined by end-users rather than by editors. Since this type of links does not directly appear in the web-pages, we called them "implicit links." For example, when a user visits a web-page, an implicit link from the user to the web-page can be formed. Another type of implicit links can be added among the web-pages when these web-pages are sequentially visited by a user. DirectHit algorithm [12] is an example which utilizes the implicit links imagined by end-users for Web search. The underlying assumption of implicit link analysis is that the more frequently a web-page is visited by users, the more important the web-page is. DirectHit represents the users' view from their interactions with the Web. Its performance depends on the quantities of the interactions.

A problem arises when we try to propose a method to combine these two types of link analysis algorithms together to improve the search precision. Intuitively, a simple way is to interpolate the "importance" output by the two algorithms. However, this solution does not fully utilize the relationship between the users and the web-pages. Since these two algorithms calculate the importance of web-pages separately, they do not reinforce each other by their calculated results. Hence, a more unified framework for link analysis is proposed in this paper. There are two assumptions in our proposed approach. The first assumption is that the importance of a web-page is influenced not only by the link structure of the web-pages but also by its visit frequency. Similarly, the importance of a user is influenced in two aspects, i.e. the link structure of the users and the relationship between

users and web-pages (which can be obtained from the users' browsing logs). The second assumption is that the more frequently a web-page is visited by "important users", the more important the web-page is. The more important web-pages a user visits, the more important the user is. In other word, the links between different users and different web-pages are not equally important and their importance is weighted by the importance of the web-page and the user. Furthermore, the two types of nodes (page nodes and user nodes) can reinforce each other. Our proposed approach can be used to analyze the access log of a website to judge the importance of the web-pages and mine the potential users of the website. Through experiments, we found that by using our approach, the search precision is improved by 11.8% and 25.3% compared to the traditional HITS and DirectHit algorithm respectively.

The rest of this paper is organized as follows. In Section 2, we present related work on Web link structure analysis, including explicit link analysis and implicit link analysis. In Section 3, we present the proposed unified framework for Web link analysis, which can support both Kleinberg's HTIS algorithm and the DirectHit algorithm. Then, we show the experimental results of the system in Section 4. Finally, we conclude in Section 5.

2. Related Works

Kleinberg's HITS [1] algorithm (based on hub and authority calculation) and Google's PageRank [7] algorithm are two representative algorithms on link structure analysis. Kleinberg analyzed hyperlinks among web-pages and drew two conclusions: Firstly, different web-pages are not of equal importance, authoritative pages and hub pages are more valuable. Secondly, authoritative pages and hub pages generally reinforce each other. So, to offer better search results for user's query, we should focus on the most authoritative pages. Based on the second conclusion, Kleinberg presented the Hyperlinked-Induced Topic Search (HITS) algorithm as the following steps. (1) Use an ordinary search engine, like AltaVista, to form the root set as the starting point; (2) Get the base set by adding pages pointing to or pointed by root pages; (3) Count the authority and hub weights of each page in the base set with an iterative algorithm which can be described as follows.

For each page, let $a(p)$ and $h(p)$ denote its authority weight and hub weight, respectively, which can be calculated as below.

$$a(p) = \sum_{q \rightarrow p} h(q) \text{ and } h(p) = \sum_{p \rightarrow q} a(q)$$

Let $A=[a_{ij}]$ denote the adjacent matrix of the base set: $a_{ij}=1$ if page i has link to page j , and 0 otherwise. Vectors

\mathbf{a} and \mathbf{h} correspond to the authority and hub scores of all pages, hence,

$$\mathbf{a} = A^T \mathbf{h} \text{ and } \mathbf{h} = A \mathbf{a}$$

Actually, hub and authority scores can be obtained by calculating the eigenvector of the matrix AA^T and $A^T A$. The CLEVER system [8] of IBM Almaden Research Center implemented Kleinberg's idea. It achieves comparable performance with Yahoo! which maintains manual compilation of net resources.

Google has used the approach named PageRank [7] to evaluate the importance of web-pages. It is partially similar to the Kleinberg's authority idea and focus on the citations of a given page. This gives some approximation of the page's importance or quality. There are some difference between PageRank and Kleinberg's algorithm. First, the value of a_{ij} and a_{ji} in adjacent matrix A is normalized by the total number of out-links and in-links. Thus, a probability transition matrix M is constructed. Second, PageRank proposed a random walk to simulate a web surfer who at each time step is at some web-pages, and decided which page to visit on the next step as follows. With probability $1-\epsilon$, the user randomly picks one of the hyperlinks on the current page and jumps to the page it links to; with probability ϵ , the user "resets" by jumping to a web page picked uniformly and at random from the collection. This defines a Markov chain on the web pages, with the transition matrix $\epsilon U + (1-\epsilon)M$, where U is the transition matrix of uniform transition probabilities ($u_{ij} = 1/n$ for all i, j). The vector of PageRank scores p is then defined to be the stationary distribution satisfying $(\epsilon U + (1-\epsilon)M)^T p = p$. By applying this technique, Google gets much better results than other text-based search engines.

Because the authority idea exploits the structure information, it provides a new and simple way to understand the features of the Web. Since then, many researchers have extended the algorithms to improve their efficiency. Chakrabarti et al. [2][3] pointed out that text surround hyperlinks in source web-pages is helpful to express the content of destination web-pages. Moreover, to reduce weight factors of hyperlinks from the same domain, the problem that a single website dominates the computation can be avoided. Ng et al. [5] present ed randomized HITS and subspace HITS algorithms to enhance the stability of the basic HITS. The former imitates a random walk on web-pages and defines the authority/hub weight as a chance of visiting that page on time step t (t is large enough). The latter uses the first k eigenvectors instead of all of matrix $A^T A$ to count the authority values. Cohn et al. [6] introduced probabilistic factor into HITS and apply EM model. All these

progresses show that the authority idea has great potentials in application fields.

Besides web-pages' domain, link analysis can also be applied to users' domain to find the relationship of human beings. Graph theory has been used in social research for many years to analyze individuals' relationships in different cases [9]. A good example comes from the telephone billing graph. By searching connected and isolated components, scientists can estimate the diameter of the whole graph and go hunting for the complete sub-graph, which is referred to as clique, to indicate contacts among people. In the experiment, phone numbers are nodes/vertices and phone calls are edges. Another more interesting and direct graph is that people become nodes. In sociology, there spreads a famous phrase "six degree of separation", which became popular by a 1990 same-titled play. This means that any pair of people on the earth can get acquaintance by no more than six intermediaries. Some sociologists have made a package-delivery experiment between two cities, which partially proved the hypothesis but far from the ultimate perfect conclusion. On the contrary, some certain sub-graphs can be explored more easily and thoroughly. For instance, members of an enterprise can form a collaboration/co-operation graph. By recognizing the functional scale of each employee, the system can learn knowledge about each employee, e.g., whether he/she is an active or lonely role and thus can help make decisions.

The link analysis is expected to work on homogenous data as well as heterogeneous data. For example, the links within the same domain such as web-pages (users) are considered homogenous, and the implicit links between the users and web-pages constructed from user access logs are heterogeneous because they represent relationship of different domains (web-pages and users). A good question to ask is whether introducing heterogeneous data can improve the link analysis of the homogenous data. DirecHit [12] is one of the representative examples, which Harnesses millions of human decisions by millions of daily Internet searchers to provide more relevant and better organized search results. DirectHit's site ranking system, which is based on the concepts of "click popularity" and "stickiness," is currently used by Lycos, Hotbot, MSN, Infospace, About.com and roughly 20 other search engines. The underlying assumption is that most relevant pages of a topic are those most visited ones. Miller [13] proposed a modified HITS algorithm which utilizes the users' behaviors on the web-pages to improve the calculation of hub and authority scores. In the modified algorithm, the adjacent matrix A is replaced by a modified matrix A' to improve the performance. First, A' is initialized to A . Then, the value of a_{ij} is increased every time when a user travels from web-page i to web-

page j (these data can be obtained from the access log of a web-site). Then, the updated adjacent matrix A' is replaced to improve the authority/hub calculation. This algorithm is very similar to Google's PageRank. In PageRank, random walk is proposed to simulate the user's random click from one page to another page. In this algorithm, real users' behaviors are used to set the probability of walking from one page to another page instead of the uniform distribution in PageRank algorithm. However, this algorithm only converts heterogeneous links (links between users and web-pages) to homogenous links (links within web-pages) for calculation. The links from the users to web-pages are only used to enhance the link analysis for web-pages. And the users' importance is ignored in this algorithm. Therefore, in this paper we propose a new approach to analyzing the homogenous and heterogeneous links in a unified framework which has the following properties:

The importance of homogenous nodes can be obtained by analyzing the links within the homogenous data.

The importance of homogenous nodes can also be obtained by analyzing the links within heterogeneous data.

The importance of heterogeneous nodes can reinforce each other by their links, which is the main contribution of our proposed approach.

3. The Unified Framework for Web Link Analysis

There are clear similarities between the ideas of the Kleinberg's authoritative web-pages, the social graph, and the DirectHit algorithm: all have netlike and hierarchical structures; all nodes have clear contents and intensions but some are more authoritative than others; various relationships exist among nodes. Because these ideas have achieved big successes in their individual field, our proposal is to design a new net model which can integrate them and exert advantages of them.

Our approach is to construct a unified framework to calculate the importance of homogenous and heterogeneous data at the same time. To simplify the framework, we only deal with two sets of heterogeneous nodes in this paper. These two sets of nodes are denoted as S and T , as shown in Eq. (1).

$$\begin{cases} S = [s_1, s_2, \dots, s_n] \\ T = [t_1, t_2, \dots, t_m] \end{cases} \quad (1)$$

where, $s_i (i = 1, \dots, n)$ is a node in the homogenous node set S ; and $t_j (j = 1, \dots, m)$ is a node in the homogenous node set T . A node set is heterogeneous if it contains nodes from both S and T .

Besides the two sets of nodes, there are also two kinds of links in our proposed framework: one is homogenous link, which links two homogenous nodes, e.g. links within the set S (or T); another is heterogeneous link which links heterogeneous nodes, e.g. links between a S node and a T node. The adjacent matrixes are used to represent these two kinds of links. L_S and L_T stands for the adjacent matrixes of link structures within set S and T , respectively. L_{ST} and L_{TS} stand for the adjacent matrixes of links from S nodes to T nodes, respectively. $L_{ST}(i, j) = 1$ if there is a link from node s_i to node t_j .

There are two levels of calculations in our proposed hub/authority calculation approach: one is that the hub value and authority value of homogenous nodes reinforce each other by the homogenous links; and the other is that the importance of heterogeneous nodes reinforces each other by the heterogeneous links. The calculations in this approach are written as follows.

$$\left\{ \begin{array}{l} a(S) = \beta L_S^T h(S) + (1 - \beta) L_{ST} i(T) \\ h(S) = \beta L_S a(S) + (1 - \beta) L_{ST} i(T) \\ i(S) = a(S) + h(S) \\ \\ a(T) = \gamma L_T^T h(T) + (1 - \gamma) L_{TS} i(S) \\ h(T) = \gamma L_T a(T) + (1 - \gamma) L_{TS} i(S) \\ i(T) = a(T) + h(T) \end{array} \right. \quad (2)$$

where $a(S)$ is the authority value of the node within S , and $h(S)$ is the hub value of the node of S . Similarly, $a(T)$ and $h(T)$ stand for the authority and hub value of the node in T , respectively. $i(S)$ and $i(T)$ stand for the importance of the node in S and T , respectively. β and γ are the weight parameters to adjust the importance of homogeneous links and heterogeneous links, respectively.

At the beginning of the calculation, all vectors, $a(S)$, $h(S)$, $a(T)$ and $h(T)$ are initialized to 1. The hub values and authority values are updated using Eq. (2) iteratively. The resulting vectors at each iteration are normalized before next iteration of calculation.

The proposed approach can be easily applied to many domains, e.g. the Web environment. We have built a hybrid net model for the Web, in which users and web-pages are the nodes, and links between page-to-page, user-to-page and user-to-user are the edges. Figure 1 explicates these relations.

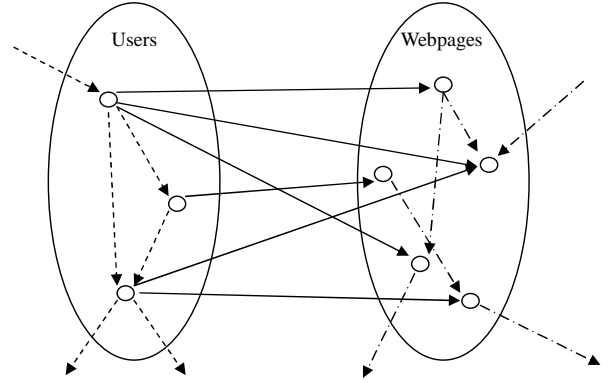


Figure 1. Relations in and between users and web-pages

Three kinds of arrows have been employed in Figure 1 to depict different kinds of links. In the web-page set, arrows from one node to another indicate hyperlinks between these pages. In the user set, arrows mean social relations of users, such as teacher-student, manager-employee etc. Arrows from users to web-pages show user's visiting actions toward web-pages which express a kind of evaluation of web-pages by users such that the authority/hub scores of a webpage can be more creditable. Because the influence of web-pages to users is obscure and hard to formalize, it is omitted in our implementation in this paper.

Given a group of users named *user set*, all web-pages that *user set* has linked form *webpage set*. For each user i , the weight u_i is given to denote the knowledge level of the user i . The higher u_i is, the more knowledgeable the user i is. For each web-page j , the authority weight a_j and the hub weight h_j are defined as done in the Kleinberg's HITS algorithm. The hub and authority values can be iteratively updated using Eq. (3).

$$\left\{ \begin{array}{l} a(p) = \sum_{q \rightarrow p} h(q) + \sum_{r \rightarrow p} u(r) \\ h(p) = \sum_{p \rightarrow q} a(q) + \sum_{r \rightarrow p} u(r) \\ u(r) = \sum_{r \rightarrow p} a(p) + \sum_{r \rightarrow q} h(q) \end{array} \right. \quad (3)$$

We can also rewrite these formulas with vectors and matrices. Let $V = [v_{ij}]$ denote the visit matrix of *user set* to *webpage set*, $v_{ij} = 1$ if user i visits page j , and 0 otherwise. Thus, the calculation of a , h , and u can be rewritten in Eq. (4), which is a simplified formula of Eq. (2).

$$\begin{cases} a = \beta A^T h + (1 - \beta) V^T u \\ h = \beta A a + (1 - \beta) V^T u \\ u = (1 - \beta) V (a + h) \end{cases} \quad (4)$$

where a is the authoritative value of a webpage, h is the hub value of a webpage, and u is the importance of a user. β is the weight parameter to adjust the influence of the Kleinberg algorithm and DirectHit algorithm. If $\beta = 1$, our algorithm is the pure Kleinberg algorithm. If $\beta = 0$, our algorithm becomes the pure DirectHit algorithm. We hope that we can find out the best β to maximize the search performance.

4. Experiments

In order to test the effectiveness of our proposed unified algorithm, we evaluate the performance of searching documents to the user's queries. Commonly, precision and recall of the search results are selected as the evaluation criteria. Since our algorithm incorporate user's visit information in calculating the authority/hub score of the web-pages, a proxy log or a search engine log is the suitable data source to obtain the users' access information. In our experiment, we use 4 days log from a proxy server at Microsoft. The log records user visit information, in which one record is corresponding to one http request for a web object from an IP address. In other words, the different users from the same IP address are considered the same user in our experiments. Some heuristic rules (e.g. the words within the hyperlinks, the extension of the filenames, etc.) are applied to filter out the unrelated information, e.g. ads, images, etc. Only text pages are reserved in the final dataset, which contains 866108 visit records to 710005 pages by 27388 users.

In our experiment, the connection graph of web-pages for a given query is constructed by a way similar to HITS algorithm. That is, the query is first sent to a text-based search engine, i.e. AltaVista, and the top 200 web-pages matching the query are retained as the *root set*. Then, the *root set* is expanded to the *base set* by its *neighborhood*, which is the set of web-pages that either point to or are pointed to by pages in the *root set*. In practice, we set the maximum in-degree of nodes as 50, which is commonly adopted by the previous work [1, 4]. The expanded set of web-pages forms the nodes of the *neighborhood graph*. Hyperlinks between web-pages not on the same web site form the directed edges. In this way we construct the matrix A in our algorithm. In the *root set* there will be a number of web-pages that have been visited previously by other users, thus being recorded in the log. The visits to the overlapped web-pages form the directed edges in the *user-web page graph*, and form the non-zero entries in the

matrix V , whose rows represent different users and columns represent web-pages. Since we only use 4 days proxy log to construct the user-web page graph, the visit information is very sparse. To avoid too many zeros in matrix V , we smooth the non-visited entries in matrix V by some very small numbers, for example, if total number of zero entries is N , those zero entries can be replaced by $1/N$, representing the probability of a random visit to the page. The percentage of the overlapped pages is determined by the specific query and the dataset. If none of the web-pages in the *root set* is recorded in the log, the algorithm will essentially be degraded to the HITS algorithm.

Since 4 days proxy log is not enough to simulate the users' view on the Web, in our experiments we select the queries which are popular topics in the accessed logs to increase the web-pages converge between the *root set* and proxy log. Obviously, the results for the queries which have no overlapped web-pages between the root set and proxy log will be the same as HITS algorithm. Finally, 10 queries are selected for evaluation, as shown in Table 1. The percentage of pages in the *root set* recorded in the log is given beside each query.

ID	Query	TN	PN	UN
1	audi car	5051	21	20
2	baby care	3010	7	50
3	windows XP	5046	21	372
4	C++ source code	3053	7	22
5	computer vision	3043	3	3
6	daily news	3017	25	170
7	notebook computer	3042	10	57
8	online dictionary	3045	9	45
9	network security	3029	13	268
10	online music	3041	16	79

Table 1. 10 queries for evaluation.

In Table 1, the abbreviation TN means total number of pages in the formed sub-graph. PN means the number of pages in the root set that can be found in the proxy log. UN means the number of different users that have visited the found pages in the proxy log.

We compare our algorithm with the pure text-based retrieval, HITS algorithm and DirectHit algorithm based on precision and relative recall at 5 and 10 documents. We run the 4 algorithms on each of the query and evaluate the precision for top 10 documents. The set of all top ranked documents from the 4 algorithms are pooled together and rated for relevance by 5 volunteers. And the final relevance judgment for each document is decided by majority votes. We then computed precision at 10 documents for each algorithm-topic pair. For HITS and our algorithm we evaluate the returned top authority pages. In those cases where the number of overlapped pages, say PN in Table 1, is less than 10, we calculate its precision

by dividing the number of positively labeled pages by PN. In the experiment, we set the parameter β of our algorithm to 0.6, and the selection of the parameter will be discussed in following section. The comparison of precision for 4 algorithms is shown in Figure 2. The result labeled as *avg1* is the average of all of the above 10 queries including the extreme case in which HITS failed totally. We also showed the average precision *avg2* which do not include the extreme case to give a more stable average precision value of the four algorithms in the small test set. According to *avg2*, we found that our proposed unified algorithm outperform the basic HITS algorithm and DirectHit algorithm. The average improvement of precision over the HITS is 11.8% and DirectHit is 25.3%.

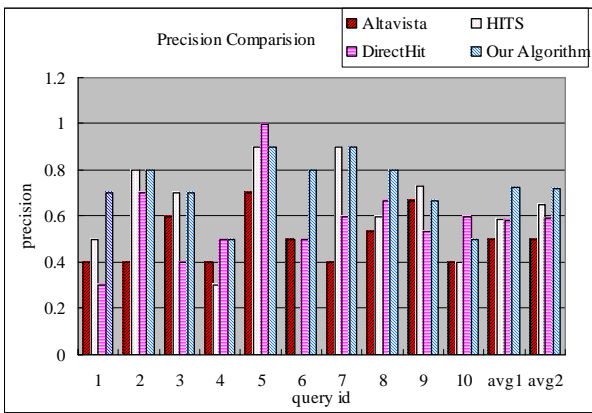


Figure 2. The precision comparison of 4 algorithms

Our experimental results show a strong trend that when the user visit information to the root set is more abundant, the search precision of our algorithm is significantly improved. For example, for the query “audi car” 21 users visited 10 pages in the root set, which is higher than other queries, leading to a significant improvement in precision compared with the other two algorithms. The returned top ranked pages for query “audi car” are listed in Table 2

From Table 2, we found that the percentage of good authority pages in top results from our algorithms is much higher. More importantly, the result of our proposed algorithm is not a simple combination of two rankings from HITS and DirectHit algorithms. Undesired URLs like “<http://www.volkswagen-car-body-parts.com/>” from HITS and “<http://www.edmunds.com/used/>” from DirectHit, which are top ranked pages in the two algorithms, are ranked lower in our algorithm. On the other hand, much more authoritative URL such as “<http://www.s-cars.org/>” is ranked much higher. This is mainly because the links from users to pages carry important information, and the reinforcement between users and pages plays an important role in the ranking of pages. A simple example to illustrate the effect of our

algorithm is that if more users visit a page i and thus increase its authority weight, the hub/authority weights of the pages it points to or point to it are also changed. As a result, the importance weights of users and the authority/hub weights of pages are modified iteratively through mutual reinforcement.

	HITS algorithm	DirectHit	Our proposed algorithm
1	http://www.quattroclubusa.org/	http://www.audiusa.com/	http://www.audiusa.com/
2	http://www.audiworld.com/	http://pages.ebay.com/ebaymotors/browse/cars.html	http://www.audiworld.com/
3	http://www.volkswagen-car-body-parts.com/	http://www.audiworld.com/	http://www.quattroclubusa.org/
4	http://www.auto-body-parts-zone.com/	http://www.edmunds.com/used/	http://www.vwvortex.com/
5	http://www.car-parts-car-body-parts.com/	http://www.nytimes.com/pages/automobiles/index.html	http://www.s-cars.org/
6	http://www.s-cars.org/	http://www.autotrader.com/	http://www.audi.co.za/
7	http://www.vwvortex.com/	http://www.thecarconnection.com/	http://www.a4.org/
8	http://www.audiusa.com/	http://www.uvas.com/	http://www.porsche.com
9	http://www.honda-auto-body-parts.com/	http://www.vwvortex.com/	http://www.audifans.com/
10	http://www.mazda-car-body-parts.com/	http://www.gearheadcafe.com/mags.html	http://www.karquattro.com/

Table 2. Top 10 results from the three algorithms for query “audi car”

We also found that sometimes the precision is not quite sufficient for evaluation. In other words, binary judgment to a web-page can not fully reflect the value of a web-page. The precision for some queries are not improved a lot compared to HITS, but we found that more authoritative pages are ranked higher. The top 10 results for the query “baby care” shown in Table 3 is a good example of this.

From Table 3, we found that although the precision of our proposed algorithm is similar to the HITS algorithm, our results are more reasonable. For example, the authoritative URL “<http://www.thebaby.net/>” obtains higher ranking in our proposed algorithm.

In some cases our results were biased compared with HITS when the roots set returned by AltaVista contained

URLs that are extremely popular or correspond to portals which are not relevant to any queries. The visits to those sites are extremely frequent because users often begin their search/browse from those sites. The results will be biased because of the unexpected high score of those sites. One way to solve the problem is to remove those sites heuristically, or set the visit links to those abnormally popular sites with a small number thus reduce its influence to other sites.

	HITS algorithm	DirectHit	Our proposed algorithm
1	http://www.parentsoup.com	http://www.thebabylane.com	http://www.thebabylane.com
2	http://www.parentsplace.com	http://www.webmd.com	http://www.parentsoup.com
3	http://www.pampers.com	http://www.babyplace.com	http://www.parentsplace.com
4	http://www.tnpc.com	http://www.thebabycorner.com	http://www.pampers.com
5	http://www.noah-health.org/english/pregnancy/pregnancy.html	http://www.amazingbaby.com	http://www.thebaby.net
6	http://www.baby-care.com	http://www.playtextbaby.com	http://www.babybag.com
7	http://www.peapods.com	http://www.wellbeing.com	http://www.tnpc.com
8	http://www.yourbaby.com		http://www.beechnut.com
9	http://www.thebaby.net		http://babycatalog.com
10	http://www.beechnut.com		http://www.peapods.com

Table 3. Top 10 results from the three algorithms for query “baby care.”

We also found in our experiments that in some extreme situations, the HITS algorithm failed to generate good results when a group of very similar sites using different host names and all link to each other thus they reinforce each other and got un-expected high score. In these cases the results of HITS will be seriously biased from the query topic, e.g. the query of “daily news”, as shown in Table 4.

As can be seen from Table 4, HITS failed to produce good results because of the strong reinforcement between a group of pages. Since in our algorithm, the visit links are made by the Web users instead of the Web editors, the

returned results will be re-ranked more reasonable than HITS algorithm.

	<u>HITS</u>
1	http://gamespy.pricegrabber.com/
2	http://www.gamespy.com/software/
3	http://www.gamespy.com/network/pc.shtm
4	http://www.gamespy.com/network/console.shtm
5	http://www.gamespydaily.com/
6	http://www.fileplanet.com/
7	http://www.gamespy.com/
8	http://gamespy.pricegrabber.com
9	http://www.planetdeusex.com/
10	http://www.gamespy3d.com/
11	http://www.strategyplanet.com/
12	http://www.sportplanet.com/
13	http://www.gamespyarcade.com/
14	http://www.planetps2.com/
15	http://www.planetxbox.com/
16	http://www.classicgaming.com/
17	http://www.rpgplanet.com/
18	http://www.planetdreamcast.com/
19	http://www.planetunreal.com/
20	http://www.planetfortress.com/
	<u>our algorithm</u>
1	http://www.news.com/
2	http://news.bbc.co.uk/
3	http://www.cnn.com/
4	http://www.latimes.com/
5	http://www.google.com/
6	http://www.drudgereport.com/
7	http://www.lycos.com/
8	http://www.newslinx.com/
9	http://www.washtimes.com/
10	http://www.businessweek.com/
11	http://www.computernewsdaily.com/
12	http://gamespy.pricegrabber.com/
13	http://www.gamespydaily.com/
14	http://www.gamespy.com/software/
15	http://www.gamespy.com/network/pc.shtm
16	http://www.gamespy.com/network/console.shtm
17	http://www.fileplanet.com/
18	http://www.gamespy.com/
19	http://www.1stheadlines.com/
20	http://www.sciencedaily.com/

Table 4. Top 20 results of query “daily news” by HITS and our proposed algorithm

Since we modify the equation of the HITS algorithm, it is difficult to prove that our proposed algorithm will always converge after several iterations. Hence, we perform an experiment to test the convergence of our

proposed algorithm. The difference of the page authority and hub scores and user weight is plotted for each iteration, ranging from 1 to 20 iteration. The difference d is defined as $d = \sum (w_i - w_{i-1})^2$, where w_i represents the authority/hub values at iteration i .

From Figure 3 we found that the difference of the authority/hub values between two consecutive iterations drops significantly after 5 iterations and shows a strong tendency to zero. In our experiments, we found that the algorithm converges with different values of β . This proves the convergence of our algorithm in a practical way.

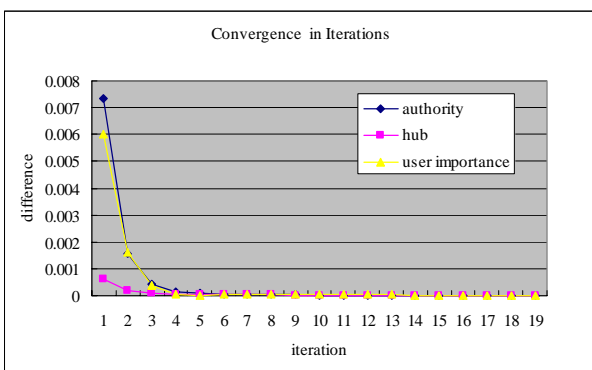


Figure 3. The converge of our proposed algorithm ($\beta=0.4$)

The parameter β in our algorithm indicates the contribution of two different kinds of links in calculating the hub/authority score. When β is set to its boundary value, say 0 or 1, our algorithm will degrade to HITS or DirectHit respectively. We hope that parameter β can be optimized to a specific value so that our algorithm can get the best performance. We run our algorithm with parameter β increasing from 0 to 1 with the step of 0.2 based on the 10 selected queries and calculated the precision for top 10 documents in a similar way stated above. The results do not show much difference in precision, probably due to the small log set we are using. But the ranking of pages do varies with different parameter β , more experiments will be conducted in the future to test the influence of parameter β and how to choose the optimal value.

5. Conclusions

In this paper, we proposed a unified framework for Web link analysis. First, the hyperlinks embedded in the web-pages and the interactions of the users with the web-pages can be analyzed in this unified framework. The Kleinberg's HITS algorithm and DirectHit algorithm can be considered two instances of our proposed framework. The importance of the web-pages and users can reinforce

each other. The resulting web-page importance was used to re-rank the retrieved results and showed that the search performance can be significantly improved by our proposed approach. Furthermore, our proposed algorithm can also mine the potential users of some web pages.

6. References

- [1] Jon Kleinberg, Authoritative sources in a hyperlinked environment, in: Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [2] Soumen Chakrabarti et al., Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, in: Proc. of the 7th International World Wide Web Conference, 1998.
- [3] Soumen Chakrabarti et al., Mining the Link Structure of the World Wide Web, IEEE Computer Vol.32 No.8, Aug 1999.
- [4] K Bharat and M Henzinger, Improved Algorithms for Topic Distillation in a Hyperlinked Environment, in: Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.
- [5] Andrew Y. Ng et al., Stable algorithms for link analysis, in: Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.
- [6] David Cohn et al., Learning to Probabilistically Identify Authoritative Documents, in: Proc. of the 17th International Conference on Machine Learning, 2000.
- [7] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in: Proc. of the 7th international World Wide Web Conference Vol.7, 1998
- [8] IBM Almaden Research Center, The Clever Searching, the Clever project of IBM Almaden Research Center.
- [9] Brian Hayes, Graph Theory in Practice, American Scientist, Jan. to Feb. 2000
- [10] Garry McGovern, Predictions for 2002, New Thinking, Dec 31, 2001
- [11] T. Berners-Lee, J. Handler, and O. Lassila, The Semantic Web, Scientific American, May 2001.
- [12] DirecHit: <http://www.directhit.com>.
- [13] Joel C. Miller, Gregory Rae, Fred Schaefer. Modifications of Kleinberg's HITS algorithms Using Matrix Exponentiation and Web Log Records, in: Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.