XRank: Learning More from Web User Behaviors

Yi Zhang, Lei Zhang, Yan Zhang †
National Laboratory on Machine Perception
Peking University
100871 Beijing, China
{zhangyi, zhangl, zhy}@cis.pku.edu.cn

Xiaoming Li
Department of Computer Science and Technology
Peking University
100871 Beijing, China
lxm@pku.edu.cn

Abstract

Link analysis has been widely used to evaluate the importance of web pages. PageRank, the most famous link analysis algorithm, offers an effective way to rank the pages. However, the algorithm ignores three facts. First, nowadays the way that users retrieve information is quite different from the previous way when web search engine was not extensively used. Second, inter-site links and intra-site links should not be treated equally. A link from a different site is more important for a page than that within the same site. Third, most users start their browsing from a homepage, which should be given more weight than other pages. In this paper, we propose a novel ranking algorithm called XRank as a solution to these problems. Experimental results on the CWT100g show that our XRank algorithm outperforms other famous ranking algorithms, including PageRank and Two-Layer PageRank, especially on sites recommendation and web spam avoidance.

1 Introduction

The combination of relying on web search engines and following hyperlinks has become the most widely used method in retrieving information nowadays. Before search engine became the predominant tool of finding information on the web, users mostly accessed the web content via following hyperlinks. However, with the emergence of booming web pages, locating the target page is extraordinary difficult. Web search engines have gradually become an indispensable assistant in our life. Many recent studies have confirmed the growing importance of search engines [12, 7]. According to [3], web users spend a totally 13 million hours per month interacting with Google alone. Once there is a said “if your page is not indexed by Google, your page does not exist on the Web” [2]. Though the statement may be a little exaggerative, it reflects some truth indeed.

Link analysis algorithms, which convey the relative importance of web pages by exploiting the web link structure, effectively help the web search engine to evaluate the quality of the pages and rank the correlative results. There are many well known link analysis algorithms, such as PageRank [5, 13] and HITS [11, 4]. Among them, PageRank, which plays a key role in Google search engine, is of great success and widely studied.

User’s walk model is not simply random but rather highly influenced. The traditional version of PageRank [13] corresponds to the standing probability distribution of a random walk on the web graph which the web viewer simply keeps clicking on successive links at random or gets tired and jumps to another page. However, experimental results show that today the original PageRank algorithm is no longer effective because of the change of users’ surfing behaviors. At present, most web users would like to begin their browsing sessions by typing the keywords in the query box and then selecting one of the top ten results returned by the search engine [15] or by following the links of some famous homepages, such as Yahoo. For example, a user wants to get some information about Peking University, but he does not remember its exact URL. To visit the page, usually he will first go to Google search engine to issue the query “Peking University”. Then he clicks on the proper link in the result page and arrives at the Peking University homepage. Further, he reaches other pages from this homepage. However, once he is familiar with this homepage, he will directly locate it without consulting search engines. The web user’s browsing pattern is illustrated in Figure 1. From this example, we know that the visit probability of every page is certainly not equal.

Besides, when computing the PageRank, a link between two sites which is called an inter-link should be given more weight than a link within a site namely an intra-link [10, 8]. On the one hand, due to the broad use of search engines in guiding today’s web traffic, some people who are blinded by
gain aim at making their pages rank highly by playing with web page features that the search engines’ ranking algorithms base on. This phenomenon is usually called “search engine spamming” [14]. Link spamming [9], one of the famous spamming techniques, is performed by adding a large number of in-links to one page or making the pages point to each other mutually to form a spam farm. It terribly deteriorates link-based ranking algorithms and leads to bad returned results that the top 10 are mostly taken up by the pages from the same site. On the other hand, the quality of the page is closely correlative with that of the site the page belongs to. In most situations, a page with high quality is not supposed to show up on a notorious site. So we assume that the ranking of sites can reflect the quality of their pages to some extent. Therefore, the links from other sites should contribute much more than those from the same site.

According to our analysis above, we propose XRank, a novel algorithm that fully considers these three referring points. We divide the computation of XRank in three steps. First, the initial value for every site which represents its popularity is computed. Secondly, we calculate the importance of the site in terms of the site link graph. During this process, we abstract each web page’s URL on the site level and only consider the links among sites. Finally, we calculate the importance of pages, supposing the total SiteRank score is assigned to its homepage.

The paper is structured as follows. The details of XRank algorithm are introduced in Section 2. It starts with some assumptions and definitions such as SiteRank, and then elaborates each step of the algorithm. Section 3 evaluates the method based on experimental results. The related work is presented in Section 4. Finally, in Section 5 we conclude with a discussion and future work.

2 XRank

2.1 Assumptions and Definitions

Before computing the page importance with our algorithm, some assumptions and definitions need to be presented.

We assume that currently the user’s walk model is as follows. At the beginning of each browsing session, a user selects a site’s homepage or asks a search engine for recommendation. Then he or she will leave for another page by following the hyperlinks of the homepage or the result page returned by the search engine.

It is obvious that the prime step of accessing way is quite different from that of the random surfer model. The user’s surfing behavior is greatly influenced by the web search engine and the popularity of the site. Therefore, we need to choose a measurement to judge how much the search engines bias the web pages and how famous the site is. Next, we give two corresponding definitions.

When a user queries the web search engine, the results are ordered by the PageRank values of relevant pages. The higher the page’s value is, the more chance the page will be retrieved. At the same time, the site is possibly visited and acquires certain reputation. So we define the public popularity of a site by summing up the ranking values of all pages that belong to this site. The public popularity of one site is an index of its reputation. The higher this value is, the more likely that search engines will return it to users.

Definition 1: Public Popularity

Let $PP(s)$ be the public popularity of site $S$. $C$ is the set of the pages which belong to site $S$. Then the public popularity of the site is computed by the following equation:

$$PP(s) = \sum_{p_i \in C} PR(p_i) \quad (1)$$

where $PR(p_i)$ is the PageRank of web page $p_i$.

As discussed before, the inter-links and intra-links should be treated differently. The links which point to each other mutually within one site can surely increase the number of in-links to improve the ranking of the site. However, in reality, these links contribute to the site’s quality less than the links between sites. The recommendation from different sites is more convicitive and valuable. Also experimental results tell us that the host abstraction is the best choice for the hierarchical link aggregation [17]. Then we define SiteRank by computing the ranking value based on the site link graph.

Definition 2: SiteRank

Let $S$ be a site. Let the $F_s$ be the set of sites that $S$ points to and $B_s$ be the set of sites that point to $S$. Let $M_s = |F_s|$ be the number of inter-links from $S$ and let $c$ be a factor used for normalization. $N_{u,s}$ means the number of links that are from site $U$ to site $S$. $SR(s)$, the simple version of SiteRank of site $S$, is defined as equation (2) which is

$$SR(s) = \frac{\sum_{u \in F_s} N_{u,s}}{M_s} \quad (2)$$

The paper is structured as follows. The details of XRank algorithm are introduced in Section 2. It starts with some assumptions and definitions such as SiteRank, and then elaborates each step of the algorithm. Section 3 evaluates the method based on experimental results. The related work is presented in Section 4. Finally, in Section 5 we conclude with a discussion and future work.

Figure 1. Web User Accessing Behaviors

![Figure 1](#)
derived from the idea of original PageRank algorithm [13].

\[ SR(s) = c \sum_{u \in B_s} \frac{SR(u) \times N_{u,s}}{M_u} \]  

We discuss these two definitions in detail in the following sections.

### 2.2 Calculating Public Popularity of Sites

The web can be thought as a graph with the pages as nodes and the links as edges. Every page has several forward links (out-edges) and back links (in-edges). We use the random surfer version of PageRank algorithm [13] to rank all the pages. The more back links the page owns, the more importance the page obtains. The PageRank of a page \( A \) is defined as follows:

\[ PR(A) = \frac{\epsilon}{n} + (1 - \epsilon) \times \sum_{i=1}^{n} \frac{PR(T_i)/C(T_i)}{n} \]  

where \( \epsilon \) is a damping factor, which is usually set between 0.1 and 0.2. In our paper, we use 0.15 that is usually adopted; \( n \) is the number of nodes of the whole page link graph; \( T_i \) points to \( A \) and \( C(T_i) \) is the number of out-edges of page \( T_i \).

The ranking of the page indirectly tells us the querying probability of the page when using the web search engine. Then we calculate the public popularity of each site according to definition 1 and use it as the initial value of SiteRank in the recursion.

### 2.3 Calculating Site Importance

Given the initial ranking value of every site, computation of the SiteRank based on the site link graph is available. We represent the site link graph as a matrix. Only the links between different sites are considered. We neglect the links within a site when producing the site link matrix. Then we implement the algorithm by using the matrix.

Originally thought expressed in equation (2) is that the SiteRank for site \( S \) can be computed by summing up the importance of the other sites that point to \( S \). In practice, many sites have no in-links from other sites, so the above equation is modified in the same way of PageRank [13] to obtain a random walk model to deal with the issue. When user begins browsing, he or she prefers to visit the homepage of a site judging by the site’s popularity in his or her mind. Then with the probability \( 1 - \epsilon \), a user randomly chooses one of the links on the current site and jumps to another site. With the probability \( \epsilon \), the user will be “initialized” by jumping to another site uniformly from the collection of the sites. We know that every user will have his unique opinion about popularity of a same site. However, the approximate value of all the people can be represented by the site’s public popularity. Therefore, the site importance formula of site \( S \) is modified as equation (4) similar to that of PageRank [13] where we use the site’s public popularity as the initial SiteRank value.

\[ SR(s) = \epsilon \frac{1}{n} + (1 - \epsilon) \times \sum_{u \in B_s} \frac{SR(u) \times N_{u,s}}{M_u} \]  

where \( \epsilon \) is a damping factor between sites; \( n \) is the number of nodes in the site link graph; The meanings of other variables have been discussed in definition 2.

### 2.4 Calculating Page Importance

After calculating the site importance called SiteRank, we propose the method to compute the importance of pages. Because users mostly begin their browsing session from the homepage of some site, the SiteRank can indicate the visit probability of the homepage. Subsequently, the users visit pages by following hyperlinks. Therefore, the importance is diffused among pages. So in our iterative calculation, we set each SiteRank as the initial importance value of its homepage and set other pages’ initial values zero. Then the importance of each page is calculated according to the importance of the homepage and the link structure of the whole web. Suppose that all the homepages of all sites have been crawled, we use a formula similar to the original PageRank algorithm [13] to calculate this score recursively except for setting page’s initial importance discriminately. Actually, our method, which is called XRank, is a personalized PageRank algorithm [13].

### 3 Experiments

#### 3.1 The CWT100g Data Set

We conduct our experiments on the CWT100g Data Set, the biggest Chinese web test collection with 100GB web pages [1]. The collection of CWT100g is composed of three parts: the documents, the queries and relevance judgements. Only the documents of the data set are used in our experiment. The documents are according to the statistic of Tianwang Search Engine on Feb.1, 2004. Every page in the collection has a “text/html” or “text/plain” MIME type received from the HTTP server response message. After cleaning the data set, the collection consists of 4,865,066 web pages from 17,760 sites. There are totally 477,173 inter-links and 36,753,069 intra-links in the collection.

#### 3.2 Evaluation Methods

In order to measure the retrieval performance of different ranking algorithms, we evaluate them in two ways. First, we
Table 1. Top 100 URLs of PageRank

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>drams.anhuinews.com/ahnews/includes/foot.html</td>
</tr>
<tr>
<td>2</td>
<td>et.21cn.com</td>
</tr>
<tr>
<td>3</td>
<td>drams.anhuinews.com/ahnews/includes/guidex.shtml</td>
</tr>
<tr>
<td>4</td>
<td>et.21cn.com/portray/gangtai/2002-10-10/794444.html</td>
</tr>
<tr>
<td>5</td>
<td>mms.sohu.com</td>
</tr>
<tr>
<td>6</td>
<td>antivirus.pchome.net/list.html</td>
</tr>
<tr>
<td>7</td>
<td>csonline.com.cn/lg/cs_top.htm</td>
</tr>
<tr>
<td>8</td>
<td>csonline.com.cn/lg/cs_d.htm</td>
</tr>
<tr>
<td>9</td>
<td>chinabyte.com/Gameall/731834933/944770560/index.shtml</td>
</tr>
<tr>
<td>10</td>
<td>drams.anhuinews.com/ahnews/photo/index.htm</td>
</tr>
<tr>
<td>11</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 2. Top 100 URLs of TL PageRank

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>et.21cn.com</td>
</tr>
<tr>
<td>2</td>
<td>et.21cn.com/portray/gangtai/2002-10-10/794444.html</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.phoenixtv.com/home/English/index657.html">www.phoenixtv.com/home/English/index657.html</a></td>
</tr>
<tr>
<td>4</td>
<td>et.21cn.com/wallpaper/501/index.shtml</td>
</tr>
<tr>
<td>5</td>
<td>et.21cn.com/target502.html</td>
</tr>
<tr>
<td>6</td>
<td>et.21cn.com/topic/js/ettjhg.html</td>
</tr>
<tr>
<td>7</td>
<td>et.21cn.com/love/qingse/2003/11/19/1347023.shtml</td>
</tr>
<tr>
<td>8</td>
<td>et.21cn.com/topic/star/sex/index.htm</td>
</tr>
<tr>
<td>9</td>
<td>et.21cn.com/topic/beat/1229/index2.htm</td>
</tr>
<tr>
<td>10</td>
<td>et.21cn.com/topic/star/mm/index.htm</td>
</tr>
<tr>
<td>11</td>
<td>...</td>
</tr>
</tbody>
</table>

compare the top 100 URLs returned by XRank algorithm and by other two well-known algorithms, traditional PageRank and Two-layer PageRank. Then we count the number of different sites in the top 100 URLs set of the three algorithms from which we can learn the capability of every method to recommend sites. Also, we discuss how much spam pages will affect the ranking results in the three algorithms. Details are shown in the following parts.

3.3 Participant Algorithms

We describe two famous ranking methods to compare with our proposed XRank algorithm.

3.3.1 PageRank

We implement the traditional PageRank [5, 13] algorithm and apply it to the link matrix from the link structure of the CWT100g Data Set.

3.3.2 Two-Layer PageRank

We use the two-layer PageRank algorithm [16] to rank the pages by dividing the computation into two parts. The PageRank is calculated on the host-layer and page-layer inside a host separately. Then, we apply a weighted product between the two ranking values to obtain the final global ranking for all the web pages. In our experiment, we consider that the PageRank values on two layers are of equivalent importance and make a simple product between them as the ranking foundation.

3.3.3 XRank

XRank is our proposed algorithm that considers more about the web user behaviors nowadays. We implement the algorithm in three steps as mentioned in Section 2. To make it more practical, we modify the algorithm slightly. For the CWT100g Data Set, not every homepage has been crawled so that we can not set SiteRank of each site to its corresponding homepage when computing page importance. Therefore, experimentally, we adjust our algorithm a little. If the homepage of the site is collected, we directly set the SiteRank to the initial importance of its homepage and set zero to the initial value of other pages within the same site. Otherwise, if the homepage has not been crawled, we divide the SiteRank of this site proportionately into parts which are given to each page of the site as the initial importance.

3.4 Experimental Results

Several parameters for our experiments are given ahead. During the computation of PageRank and SiteRank, we set the damping factor $\epsilon = 0.15$ and the iterative precision $10^{-4}$. We suppose that a web user browses randomly by following the links of the current page or site with the probability 0.85 and uniformly jumps to another page or site with the probability 0.15. In addition, if the distance between two adjacent iterative values for each page or site is equal or smaller than $10^{-4}$, the calculation process will terminate.

3.4.1 Top 100 Ranking URLs

We take the top 100 ranking URLs of each algorithm to compare their retrieval performance. The lists are shown in
Table 3. Top 100 URLs of XRank

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.zgjccbs.com">www.zgjccbs.com</a></td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.tp-lawbook.com">www.tp-lawbook.com</a></td>
</tr>
<tr>
<td>3</td>
<td>ko.sohu.com</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://www.cnbaolian.com">www.cnbaolian.com</a></td>
</tr>
<tr>
<td>5</td>
<td><a href="http://www.asiabt.com">www.asiabt.com</a></td>
</tr>
<tr>
<td>6</td>
<td><a href="http://www.our80s.com">www.our80s.com</a></td>
</tr>
<tr>
<td>7</td>
<td>0570.fecom.cn</td>
</tr>
<tr>
<td>8</td>
<td><a href="http://www.3cb2b.com">www.3cb2b.com</a></td>
</tr>
<tr>
<td>9</td>
<td>guangzhou.cnfortune.net</td>
</tr>
<tr>
<td>10</td>
<td>abroad.cnfortune.net</td>
</tr>
<tr>
<td>11</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1, Table 2 and Table 3 which are respectively in accord with traditional PageRank, Two-Layer PageRank and XRank.

As we see in the tables, XRank algorithm gives a higher performance than other algorithms on the quality and diversity of the recommendatory pages. After evaluating the quality of each ranking result through human judgements on the ordering of the top 100 URLs, we learn that 56 pages of XRank result are considered to be in the top 100 while PageRank is 39 and Two-layer PageRank is 21. In addition, the random surfer version of PageRank strongly holds the idea that the number of in-links decides the importance of the page. More than 380 thousand pages are from the same site, drams.anhuinews.com, so that the pages of the site improve their ranks a lot by mutually pointing to each other to form a link farm. For Two-Layer PageRank, the weight is hard to balance when making the product of the PageRank values between the two different levels. So its performance is not always better than the PageRank. As shown in Table 2, the top 10 popular pages are almost dominated by the same site-et.21cn.com. This ranking method can easily lead to web spam as well as weaken new pages to become popular. Inversely, our proposed algorithm, XRank shows great improvement on the performance. As shown in Table 3, XRank contains more famous consumer-oriented sites and the top 100 URLs are mostly represented in the abstraction of site. Besides, the pages belong to diverse fields, for example, news, business, education, etc. To a certain extent, the ranking method can avoid pages from forming spam farms.

3.4.2 The Variety of Sites Recommendation

Statistical results indicate that XRank algorithm shows a stronger capability to recommend more sites than that of other two algorithms. For each ranking method, we put its top 100 ranking URLs into 10 buckets averagely according to their ranking values and organize them in a descending order. The pages in bucket 1 have the highest 10 ranking values and the pages in bucket 10 have the lowest 10 ranking values. Figure 2. shows the distribution of site number within each of the top 10 buckets. In the top 10 buckets, except for bucket 9 where Two-Layer PageRank and XRank recommend no new sites, however, totally our XRank introduces up to 61 various sites while the traditional PageRank refers only 30 and Two-Layer PageRank introduces merely 26. Besides, it is obviously seen that the ability of XRank to recommend new sites is extremely outstanding in the top 5 buckets.

4 Related Work

For years, there has been a great deal of work on link analysis. All link analysis algorithms are based on the consideration that web is treated as a graph, with the pages represented by nodes and the links between pages functioning as the edges, connecting the nodes. The number and quality of edges between nodes usually represent relevance or endorse some authority.

Kleigberg proposes a ranking algorithm of finding authoritative pages on a given topic [11]. The algorithm is known as HITS. However, the algorithm is vulnerable to the situation which the root link has few in-links but a large amount of out-links. Always the community associated with the root link will dominate the result. Moreover, because HITS is an online algorithm, it needs to be calculated repetitively once a new query word is given.

Weighted PageRank is an effective improved PageRank algorithm suggested by X.M.Jiang et al [10]. It divides the whole web graph into blocks. So the links between the different blocks and among the same blocks are assigned with different weights when scoring the web pages. But how to divide the web sets into different blocks brings a serious trouble.

Block-level PageRank is another improved algorithm which divides the whole web into different semantic blocks [6]. Even though the algorithm can evaluate the pages ob-
bjectively in different fields, it has the same problem with the Weighted PageRank which brings more difficulty in finding a good block partition.

Two-Layer PageRank algorithm is proposed by Wu et al [16]. The algorithm computes the PageRank at the host-layer and the other document-layer separately and uses their weighted product to represent the final ranking reference value. It pays much more attention on the intra-links in one site and does not consider the inter-links between sites when computing the importance of the web. Therefore, the effectiveness of this algorithm is not always credible, because it ignores the fact that one web page needs to be evaluated mainly according to the incoming links from other sites. In this paper, we compare it with our new algorithm in the experiments.

Recently, Gui-Rong Xue et al. suggests the Hierarchical Ranking algorithm which interopolates the link structure and the hierarchical structure together [17]. They partition the web space into three abstract levels: the domain level, the host level and the directory level. The importance of each page on the aggregated web is distributed to individual pages belong to the supernode using a Dissipative Heat Conductance model [18] in the hierarchical structure. The final temperature of each page gives the importance of that page. When comparing it with our XRank algorithm, XRank seems more efficient and easier to implement.

5 Conclusion

Web user behavior changes very much when search engines are becoming dominant. Considering the new characteristics from such a behavior, we introduce a new link analysis algorithm-XRank. In our work, we perform the experiment on the CWT100g. Results show that our XRank algorithm significantly improves the performance of web searching, compared with the other two algorithms-PageRank and Two-Layer PageRank.

Even though XRank algorithm achieves great improvement in information retrieval, some limitations of the algorithm need to be mentioned. From the discussions above, we acknowledge some merits of XRank, namely, it can objectively rank pages, widely recommend sites and powerfully avoid spam farms’ effect. However, we should notice that our algorithm is based on the assumption that the homepage of each site has been crawled. So in reality, to get a good ranking result, more work need to be done by crawlers to check up the whole page set to guarantee each homepage is collected. In addition, our algorithm computes SiteRank according to the site link graph which only considers the inter-links between sites. However, it is usually seen that some well-known sites rarely link with each other on the economic consideration. They would not like to point to other sites of the same field and share customers. This re-ally brings us obstacles to evaluate the ranking of each page impersonally. Nevertheless, since most sites are built in a good connective group, XRank algorithm will be of great effectivity and success. In future work, we will improve our XRank to effectively deal with more exceptions, such as there are few website homepages that are downloaded by search engines.

References