Using SiteRank for Decentralized Computation of Web Document Ranking

Jie Wu and Karl Aberer

School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPF), Lausanne
1015 Lausanne, Switzerland
{jie.wu,karl.aberer}@epfl.ch

Abstract. The *PageRank* algorithm demonstrates the significance of the computation of document ranking of general importance or authority in Web information retrieval. However, doing a PageRank computation for the whole Web graph is both time-consuming and costly. State of the art Web crawler based search engines also suffer from the latency in retrieving a complete Web graph for the computation of PageRank. We look into the problem of computing PageRank in a decentralized and timely fashion by making use of *SiteRank* and aggregating rankings from multiple sites. A SiteRank is basically the ranking generated by applying the classical PageRank algorithm to the graph of Web sites, i.e., the Web graph at the granularity of Web sites instead of Web pages. Our empirical results show that SiteRank also follows a power-law distribution. Our experimental results demonstrate that the decomposition of global Web document ranking computation by making use of SiteRank is a very promising approach for computing global document rankings in a decentralized Web search system. In particular, by sharing SiteRank among member servers, such a search system also obtains a new means to fight link spamming.

Keywords: Web information retrieval, link structure analysis, search engine, ranking algorithm, decentralized framework

1 Introduction

Link-based rank computation is very important for Web information retrieval. Classical centralized algorithms like PageRank both time-consuming and costly for the whole Web graph. We look into the problem of rank computation in a decentralized and timely fashion by making use of SiteRank and aggregating rankings from multiple sites. We start from studying the Web graph at a higher abstraction level.

1.1 Different Abstractions for the Web Graph

Previous research work focused on the page granularity of the Web, i.e., a graph where the vertices are Web pages and the edges are links among pages. A typical result is the PageRank algorithm [8]. We propose to study the Web graph at the granularity of Web site. We call the graph at the document level the *DocGraph*, and the graph at the Web site level the *SiteGraph*. We also use the notion of *SiteLink* to designate hyperlinks among Web sites and *DocLink* for those among Web documents.

W. Nejdl and P. De Bra (Eds.): AH 2004, LNCS 3137, pp. 265–274, 2004.

Definition 1 *A SiteGraph $G_S(V_S, E_S)$ of (a part of) the Web is a graph consisting of:*

- *A set V_S of vertices, where each vertex $v_s \in V_S$ represents a Web site.*
- *A set E_S of edges, where each edge $e_s \in E_S$ is a directed SiteLink.*
- *Two mappings $o_s, t_s : E_S \rightarrow V_S$, where $o_s(e_s)$ is the originating Web site and $t_s(e_s)$ is the targeting Web site of the directed SiteLink e_s.*

Similarly, we use the notations $G_D(V_D, E_D), v_d, e_d$ for a DocGraph. We call the ranking of Web sites the SiteRank for the SiteGraph and the ranking of Web documents the *DocRank* for the DocGraph. PageRank is an example of DocRank, but DocRank can be computed in a way other than PageRank, for example, as in our approach in a decentralized fashion.

1.2 Contribution of the Work

Even though the Web site graph has been studied for applications such as identification of related hosts based on linkage and co-citation, it has not been considered in the context of ranking for search engines to the best of our knowledge. Our work explores the research possibilities in this direction, proposes insights on the potential of this approach and reports on initial results of it's implementation. More concretely, we study on how to make use of the SiteGraph and the derived SiteRank to support the derivation of rankings of Web sites and documents in the sense of general importance. Our main contributions can be summarized as follows:

1. Bringing up the idea of SiteRank to describe the general importance of Web sites in the Web. After verifying that the PageRank of our sample data set follows the well-known power-law, we find that the resulting SiteRank matches this distribution as well.
2. Evaluating the correlation between the importance of a Web site and the importance of the Web documents residing on the site. It turns out that Web documents of an important Web site tend to be more important than those of the less important sites.
3. Based on the previous observations, providing a decentralized approach for computing the global document ranking in decentralized architecture for Web and P2P search and reporting on a prototype implementation of it. As a consequence, the task of global ranking computation can be performed in a decentralized fashion and its cost is widely distributed.
4. Using a shared SiteRank is a very effective anti-rank-spamming approach for search engines that are built on our decentralized architecture. We assume all participating member servers agree on a universal SiteRank in the document rank computation which allows to exclude spamming sites more easily.

In the next section we introduce our model and the algorithm to compute SiteRank. We did several sets of experiments to evaluate the significance of this idea. We first verify that the PageRank distribution of the documents stored in our crawled data set follows a power-law. Then we try to uncover the relationship between documents' PageRank and SiteRank of the corresponding Web sites. Given the observations, we believe that making jointly use of SiteRank and PageRank is an interesting direction to

determine the global ranking of Web documents in a decentralized fashion. We show the influence of SiteRank on the computation of document rankings. Finally, after a short review on related work, we conclude our work and look into future research possibilities.

2 SiteRank and Its Distribution

A natural question of studying the Web graph at the granularity of Web sites is: are Web sites somehow comparable in the sense of general importance? We will further study the implications of it in the following sections.

2.1 Random Walks in SiteGraph

Intuitively, a random walk models a simple process of randomly navigating in the Web. In a SiteGraph, an Internet user would roam around the Web sites by following Web links. A surfer with no particular interests would choose a different site with a probability roughly specified by the ratio of links to that site and the total number of outgoing links of the current site. Based on this model similar to PageRank's random walks but at a higher abstraction level, we can derive the ranking of general importance for Web sites.

Among other advantages using SiteRank opens a possibility to fight link spamming. Using information of the SiteGraph makes it difficult for PageRank spammers to spam rank of documents by creating huge number of pages pointing to a page to be spammed. Web pages are easy and inexpensive to create, thus spamming practices have become a frequent problem and nuisance in order to deceive Internet search engines. A Web site can easily, dynamically generate a large and unbounded number of dynamic Web pages by writing a simple server-side program. As a direct result the computed ranking results by algorithms like PageRank or HITS [7] are easily polluted and users have to find ways to fight rank spamming. In contrast, it is more difficult to create huge numbers of Web sites to apply such rank spamming techniques to boost the rank of a specific Web site. Other advantages of using SiteRank are discussed in [10].

2.2 The Algorithm and How to Compute

Taking the random surfing among the Web sites as a stochastic process, its transition probability matrix M_S is generated as follows:

$$M_S(i,j) = \begin{cases} \alpha_i * h_{ij} & h_i \neq 0, s_j \in ch(s_i) \\ 0 & h_i \neq 0, s_j \notin ch(s_i) \\ \frac{1}{N_S} & h_i = 0 \end{cases} \quad (1)$$

where N_S is the total number of Web sites, $s \in V_S$, simplified from v_s is a Web page, h_s is the number of SiteLinks originating from site s, $\alpha_s = \frac{1}{h_s}$ is the probability of a random surfer's following one particular SiteLink from site s, h_{ij} is the number of SiteLinks from site i to site j, $pa(s)$ is the set of parent sites of s, i.e. those sites pointing to s, $ch(s)$ is the set of child sites of s, i.e. those sites pointed to by s.

To ensure such a matrix may not have a non-trivial Eigenvector, we apply the technique of introducing a decay factor to the original SiteGraph, by the same means as that in the Page Rank algorithm:

$$M_S = p \times M_S + \frac{1-p}{N_S} \times I \tag{2}$$

where the decay factor p is usually set to 0.85 and I is the matrix whose size is the same as that of M_S and all elements have the value of 1.

Theorem 1 *The Markov chain defined by M_S for the SiteGraph has a unique stationary probability distribution.*

Proof. Omitted. Please check [10]. □

Having this transition probability matrix for the SiteGraph, we can apply the standard Power Method for computing the principal eigenvector to obtain the ranking for the Web sites.

How to compute the SiteRank for individual member servers in the decentralized search system is another important problem since none of the servers could have the global information about the SiteGraph and the SiteLinks. The approach we adopt is similar to resource discovery in distributed networks [5]. The rudimentary idea is that member search servers exchange information of SiteGraph and SiteLinks among each other such that at a certain stage, the collected partial information about the SiteGraph can lead to a sufficiently good SiteRank result approximating the SiteRank generated by a centralized global SiteGraph. As every member server learns a non-local Web graph from arriving information from others, intentional spamming of SiteLink information by a Web site could be effectively detected when substantial mismatch is observed between information from different sources.

2.3 The Distribution of SiteRank on a Campus Web Graph

In this section we give a concrete example of the results that we obtain when computing the SiteRank values for all the Web sites of a Web graph. The evaluation presented here is made on a campus-wide Web graph, the EPFL domain which contains more than 600 independent Web sites identified by their hostnames or IP addresses. We used a Web crawler to retrieve more than 2.7 million Web documents by starting from the campus portal site and following the Web links to access all the other Web sites in this domain. Using this data set we extracted the information from the member Web sites and the SiteLinks among each other, we then applied the Power Method described above to the SiteGraph to obtain the SiteRank of them. When we generated the matrix representation of this graph, those links pointing from one local page to another local page on the same site are counted by the matrix element $M_S(i, i)$.

We draw a diagram in our technical report [10] where we display on the x axis the computed SiteRank values for the sites of the campus Web, and on the y axis we display the percentage of sites that has the particular SiteRank value. The diagram is drawn in a fashion similar to that of [9]. Both axes are displayed at a natural logarithmic

scale. Suppose $Fraction(R)$ is the fraction of Web sites having SiteRank R, one of the interesting results of our work is that we found the SiteRank distribution is yet another property of the Web graph that also follows the power-law quite well:

$$Fraction(R) \propto 1/R^{0.95} \tag{3}$$

For comparison we also applied the standard PageRank algorithm to the link structure of the EPFL DocGraph to obtain the global ranking of all the Web documents in this campus Web graph. Suppose $fraction(r)$ is the fraction of pages having PageRank r, our data set shows typical power-law properties:

$$fraction(r) \propto 1/r^{1.69} \tag{4}$$

Both Log-Log figures, which can be found in [10], are not included here due to the space limit. This result is strikingly similar to that reported in a study on the Web structure [9]. Though the exponent here is a bit lower than the value found there which is around 2.1. Two reasons might account for this disparity, the difference in the nature of the different Web data sets we use and the incomplete crawling of our campus Web.

2.4 PageRank in Relation to SiteRank

Our next set of empirical experiments was conducted for elucidating the relationship between a document's PageRank and the SiteRank of the Web site the document resides on. We want to know if the intuitive assumption that importance of Web documents and Web sites is correlated, holds and in which form.

In another Fig. in [10], we display all (PageRank, SiteRank) order pairs. We find that almost all of the 1000 top ranked documents are located at the approximately top 90 sites. Furthermore, most of the top 100 documents are located at the top 30 Web sites. It appears as if there exists actually a correlation between a page's rank value and the SiteRank of its owner. Based on the experimental results and observations above, we believe our assumptions below are very reasonable:

1. v_s is important $\Rightarrow$ many important pages belong to v_s.
2. many important pages belonging to $v_s \Rightarrow v_s$ is an important Web site w.h.p. (with high probability).

Please note that these two statements are not tautological. If these statements hold true in a general sense or even if they are only true for most instead of all of the cases, we could safely distribute the weight of a Web site to its documents, proportional to their local weights, and use these distributed page weights to approximate the global ranking of documents. In the next section, we will present our preliminary results of such an attempt which shows that this approach is actually very promising for decentralized rank computation in a distributed search system.

3 SiteRank for Decentralized DocRank Computation

We want to distribute the task of computing page ranking to a set of distributed peers each of which crawls and stores a small fraction of the Web graph. Instead of setting up

a centralized storage, indexing, link analysis system to compute the global PageRank of all documents based on the global Web graph and document link structure, we intend to have a decentralized system whose participating servers compute the global ranking of their locally crawled and stored subset of Web based on the local document link structure and the global SiteRank.

3.1 SiteRank for Computation of Global Document Ranking

To fulfill our aim, we propose a decentralized architecture for search systems. First, we need to define the *external pointing set* for each document $d \in V_D$ in the DocGraph $G_D(V_D, E_D)$ with SiteGraph $G_S(V_S, E_S)$.

Definition 2 *Assume $site(d_i)$ returns the Web site that d_i belongs to, the external pointing set for a Web document d_i is defined as a set of tuples:*

$$PS(d_i) = \{(v_{si}, n_{si}) : \exists d' \; st. \; (d', d_i) \in E_D, \; site(d') = v_{si} \; and \; not \; site(d_i) = v_{si} \\ n_{si} \; being \; the \; number \; of \; such \; d' \; of \; v_{si}\}$$

For every tuple, v_{si} is a Web site that has pages pointing to d_i and n_{si} is the number of such pages on v_{si}.

We decompose our computation of the global ranking for Web documents into three steps:

1. The computation of SiteRank. The algorithm is described above. Each Web site v_s has its SiteRank value $R_s(v_s) \in (0, 1)$.
2. The computation of the local ranking of Web documents, basically we compute the local PageRank vector r_I (I means internal links) based on the DocGraph local to the Web site. A vector of weight augmentation r_E (E means external links) is also computed for all local documents. The weight element r_E for document d_i is computed as:

$$r_E^i = \begin{cases} 0, & if \; PS(d_i) \; is \; empty \\ \sum_{(v_{si}, n_{si}) \in PS(d_i)} \frac{n_{si}}{N_{si}} R_s(v_{si}), & otherwise \end{cases}$$

   where $N_{si} = \sum_{(v_{si}, n_{si}) \in PS(d_i)} n_{si}$ and $R_s(v_{si})$ is the SiteRank value of the Web site v_{si}. A local aggregation for document weight of d_i is then computed as follows:

$$r^i = w_I r_I^i + w_E r_E^i$$

   We chose the values $(w_I, w_E) = (0.2, 0.8)$ for this local aggregation. This reflects a higher valuation of external links than internal links. One motivation for this choice is the relatively low number of links across Web sites as compared to the number of links within the same Web site.
3. The application of the ranking algebra [1] to combine both rankings to produce the final global ranking. Retraction to each document gives the final global DocRank value for the page:

$$r_G^i = r^i R_s(site(d_i))$$

3.2 Case Study

As we obtain the aggregate document ranking as described above, we evaluate the results both qualitatively and quantitatively. We performed the evaluations using the following approach: we chose two selected Web sites s_1 (sicwww.epfl.ch, the home of the computing center with 280 documents) and s_2 (the support site for SUN machines with 21685 documents), with substantially different characteristics, in particular the sizes. For those domains we computed the local internal and external rankings. We also put the EPFL portal Web server s_h (www.epfl.ch) in the collection, since this is a point where most of the other Web sites are connected to. We consider this subset of documents an excellent knowledge source for information of Web site importance.

The two ranking methods to be compared qualitatively are the global PageRank computed by using the global DocGraph for the link structure, and the aggregate DocRank computed by taking our SiteRank-based approach. We examined the top 25 of the documents belonging to s_1 or s_2 resulting from both ranking methods. In our aggregate ranking, more pages of greater importance are put in the top positions. In the global PageRank, two obviously important pages are ranked much lower than some software documentation pages in the global PageRank. We can assume that this is an effect due to the agglomerate structure of these document collections. This play a much less important role in our aggregate ranking due to the way of how the ranking is composed from local rankings using SiteRank. It demonstrates what a difference in quality we have made by using SiteRank in the computation.

For quantitative comparison of rankings we adopt the Spearman's Footrule with a weighting scheme:

$$F(r_{G0}, r_{G1}) = \sum_{i=1}^{n} w_0(i) w_1(i) |r_{G0}(i) - r_{G1}(i)| \tag{5}$$

In the formula, $r_{Gj}, j = 0, 1$ are the two ranking vectors to be compared. $r_{Gj}(i)$ is the rank of document i. Please note that the rank $r_{Gj}(i)$ is different from the computed rank weight r^i_{Gj} for a document d_i. The former is the order number of a document's place in the ranking list and can only be a positive integer, for example, if a document is the topmost one of the list, its rank is 1. The latter is the actual weight value computed by the algorithm which can only be a real number between 0 and 1.

We make this weighted customization since search engines return documents in ranking order and top ones receive generally much higher attention than documents listed later. As users mostly care about top listed documents we assign 90% of the weight to the T top-listed documents for $T < n$, i.e. $w_j(i) = \frac{0.9}{T}$ for $1 \leq i \leq T$ and $w_j(i) = \frac{0.1}{n-T}$ for $T + 1 \leq i \leq n$. When $T = n$, $w_j(i) = \frac{1}{n}$ for $1 \leq i \leq n$.

We give now the results of the quantitative comparison in Fig. 1. The figure shows the ranking distances computed using the adapted Spearman's rule of different rankings with respect to the global ranking for varying values of T. Besides the aggregate ranking we include for comparison purposes other rankings that are computed for different contexts. The "subset" ranking is the ranking obtained by selecting exactly all documents that are involved in the computation of the aggregate ranking and applying the PageRank algorithm. This ranking thus uses exactly the same information that is available to

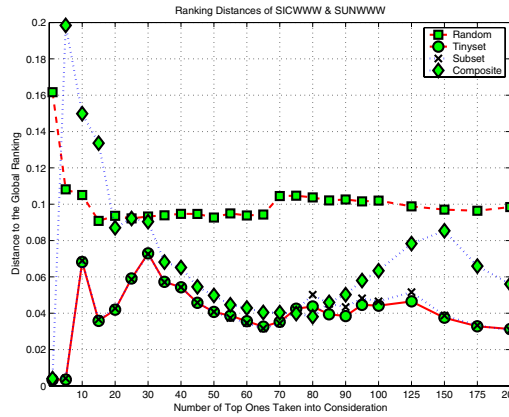

Fig. 1. Ranking Distances of SICWWW & SUNWWW

the computation of the aggregate ranking, i.e., the documents in the set $\{s_1, s_2, s_h\}$. The "tinyset" ranking is the ranking obtained by selecting exactly all documents in $\{s_1, s_2\}$ that are ranked by the aggregate ranking and applying PageRank to them. In addition, we included for calibration a randomly generated ranking. The results are shown in Fig. 1. One can observe that, interestingly, the result of the "composite" ranking appears to be much "worse" for low values of T than the global ranking. However, considering the qualitative analysis before, the result rather indicates that the global ranking seems to be poor, whereas the aggregate ranking is to be considered as the "good" ranking to be approximated. For larger values of T the aggregate ranking approximates then the rankings computed on the selected subsets. Also this is an interesting result, since the aggregate ranking is performed in a distributed manner, computing separate rankings for each of the three subdomains involved, whereas the "subset" and "tinyset" rankings can be considered as corresponding to a global ranking based on the union of the selected subdomains. This shows that by aggregation one can obtain at least as good results in a distributed manner as with global ranking using the same information.

3.3 Analysis of Reduction in Computation Cost

A member server can be a dedicated machine that crawls part of the Web. It can coexist in a Web server and compute the global document ranking for its own served Web documents. However, we need to assume that the SiteRank computation result of all Web sites in a Web graph, whose global document ranking is to be computed, is known to all member servers. This is reasonable as the number of Web sites even of whole Internet is estimated to be only at the magnitude of a dozen of million [3]. Thus the computation of the SiteRank of such a Web-scale SiteGraph is fully tractable in a low-end PC machine. Additionally, we assume that such a global SiteRank vector does not fluctuate very drastically such that it makes sense to perform such a global scale SiteRank computation infrequently and to share the result among all the member servers.

We provide a small comparison between the computation cost for the SiteRank and the PageRank. If we take the EPFL campus Web as an example, the reduction rate of the memory or disk space used to hold the matrix is:

$$\frac{Resource(SiteGraph)}{Resource(DocGraph)} = (591/2259102)^2 = 6.8 \times 10^{-6}\% \quad (6)$$

Moreover, we can use a 2-byte integer to represent every site, whereas we have to use at least a 4-byte or even 8-byte integer for pages, the rate becomes:

$$\text{reduction rate} = (25\% \sim 50\%) \times 6.8 \times 10^{-6}\% = (1.7 \sim 3.4) \times 10^{-6}\% \quad (7)$$

On the other hand, the rank computation of a matrix of size 591 can be easily performed in seconds, e.g. using a tool like Mathematica.

4 Conclusion

In many previous studies [4, 9], different snapshots of the Web have been investigated to find that not only the page in-degree, out-degree, but also the PageRank values follow the power law. We go one step further in our work to uncover that actually the SiteRank of Web sites in a Web graph also follows the power-law.

Many methods have been proposed for the rank computation of Web documents [2, 6]. However, none has been tried to decompose the computation to a two-step of first SiteRank then DocRank, which is our main contribution. One the other hand, most of existing approaches are logically centralized while ours is an inherently decentralized method. To the best of our knowledge, we are the first to use the study of the Web SiteGraph in the computation of Web document rankings for search engines in a decentralized fashion.

Based on observations on some useful correlation between the PageRank and SiteRank, we argue that decomposing the task of global Web document ranking computation to distributed participating member servers of a decentralized search system is a promising approach since we can make use of the SiteRank information to overcome the limit of a missing global view. At the same time, by doing the computation in such a complete decentralized fashion, the cost is largely reduced while we keep good quality of the ranking results. One interesting and important point is that PageRank spammers will find it difficult to spam SiteRank since they have to set up a large number of spamming Web sites to take advantage of the spamming SiteLinks.

References

1. Karl Aberer and Jie Wu. A framework for decentralized ranking in web information retrieval. In *Web Technologies and Applications: Proceedings of 5th Asia-Pacific Web Conference, APWeb 2003*, volume LNCS 2642, pages 213–226, Xi'an, China, September 2003. Springer-Verlag. September 27-29, 2003.
2. Serge Abiteboul, Mihai Preda, and Gregory Cobena. Adaptive on-line page importance computation. In *Proceedings of World Wide Wed Conference 2003 (WWW2003)*, Budapest, Hungary, May 2003. May 20-24, 2003.

3. Krishna Bharat, Bay-Wei Chang, Monika Henzinger, and Matthias Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the IEEE International Conference on Data Mining (ICDM '01)*, San Jose, USA, November 2001.
4. Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.
5. Mor Harchol-Balter, Tom Leighton, and Daniel Lewin. Resource discovery in distributed networks. In *Proceedings of the eighteenth annual ACM symposium on Principles of distributed computing*, pages 229–237. ACM Press, 1999.
6. Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Exploiting the block structure of theweb for computing pagerank. Technical report, Stanford University, March 2003. Submitted on 4th of March 2003.
7. Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.
8. Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, January 1998.
9. Gopal Pandurangan, Prabhakara Raghavan, and Eli Upfal. Using pagerank to characterize web structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.
10. Jie Wu and Karl Aberer. Using siterank in p2p information retrieval. Technical Report IC/2004/31, Swiss Federal Institute of Technology, Lausanne, Switzerland, March 2004.