Using SiteRank for P2P Web Retrieval *

Jie Wu Karl Aberer

March 24, 2004†

EPFL Technical Report ID: IC/2004/31
School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPF), Lausanne
1015 Lausanne, Switzerland
{jie.wu, karl.aberer}@epfl.ch

Abstract

Studies of the Web graph at the granularity of documents have revealed many interesting link distributions. Similarly, studies of the Web graph at the granularity of Web sites, the so-called *hostgraph*, revealed relationships among hosts based on linkage and co-citation. However, to the best of our knowledge, the graph of Web sites has not been exploited for the purpose of ranking in search engines. In this paper, we first identify the necessity of a *SiteGraph* abstraction. We derive the *SiteRank*, a ranking of general importance among the Web sites in such a graph. We then show that *SiteRank* follows a power-law distribution. As experimental data set we were using the Web of our campus with over two million documents. We uncover interesting relationships between PageRank and *SiteRank*. Based on these results and observations, we conclude 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 P2P search system. In particular, by sharing SiteRanks peers would not only be able to efficiently compute global document rankings in a decentralized manner, but also obtain a new means to fight link spamming. Our experiments give very promising results to back up the proposed ideas.

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

1 Introduction

Graph structure is a very important abstraction of the Web. The Web is hierarchically composed of entities such as domains, Web sites, and documents distributed over Web sites and linked together by hyperlinks. Note that a Web

*The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

†Written on November 14, 2003.

site corresponds not necessarily to an independent physical machine since many logical Web sites can co-reside on one physical machine. If we consider entities in the Web as the vertices and connections via links, which imply a relationship between entities, as the directed edges, the Web can be abstracted as a directed graph $G(V, E)$, where V is the set of the entities and E is the set of connections among the entities.

1.1 Different Abstractions for the Web Graph

At different abstraction levels, several such graphs can be derived to model the Web and be represented formally, e.g., as matrices. For example, when Web documents are used as the elementary entities and the links from one page to another as the directed edges, the resulting graph is the one that is widely used in algorithms like PageRank [20] and its descendants. This family of algorithms considers hyperlinks as embedded recommendation information on page importance from the viewpoint of page authors. This information is used to form a matrix representation of the Web and compute the Web document ranking for Internet search engines.

We obtain another Web graph when we abstract at a coarser granularity where we choose the individual hosts respectively Web sites as the elementary entities of the Web. The links between two hosts are the aggregation of the links between the two sets of Web pages on the hosts. In the following, we call the graph at the document level the *DocGraph*, and the graph at the Web site level the *SiteGraph*.

The *SiteGraph* was studied in earlier work [4] under the name of *hostgraph*[1], which provided several good arguments on why the abstraction at the site level is useful: documents are often represented by multiple pages and consequently it is not very reasonable to study the authorship of Web documents at the level of single pages; a Web site is usually jointly controlled and managed and may be reorganized periodically without significantly changing the semantics or linkage in relation to the rest of the Web; although very often there does not exist a direct link from local pages to a site's entry page, Web surfers can always jump there by truncating the path of the URL to navigate to the root page; statistical properties of Web pages may be skewed because of the simplicity of generating a large number of pages dynamically.

It is worth noticing that our notion of *SiteGraph* allows for the derivation of a dynamic or virtual graph of Web sites when we use dynamic or virtual relationships among Web pages instead of the static Web links. For example, when we use statistical information on navigation obtained from Web client traces, which are normally very different from the static Web link structure, as the set of edges E, we obtain a Web client trace-based *SiteGraph*. Similarly, a *DocGraph* using client traces can be defined. Thus *hostgraph* is simply one special type of *SiteGraph*s which uses the static hyper links among Web pages to define the edges. The ideas and algorithms are very easy to be applied to dynamic or virtual *SiteGraph*s which are part of our future work.

[1] We think *host* sounds more way of hardware probably because of the business of Web hosting. Thus we prefer to use *site* here to refer to the logical entity (a registered FQDN or an IP) in the Web where a Web server is running and Web requests from Web users are served. We will also see, *hostgraph* is just one special type of *SiteGraph*s we define here.

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* to support the derivation of rankings of Web sites and documents in the sense of general importance.

We first define *SiteGraph* and shortly describe random walks in such a *SiteGraph*. Then we focus on how to generate *SiteRank* [2] for a static *SiteGraph* and on how to use it in the decomposition of ranking computations. We demonstrate how this approach works by making experiments based on the data set of a crawled campus Web.

Our main contributions can be summarized as follows:

1. Identifying different types of *SiteGraph*s and their significance in the research work on Web information retrieval and Web mining;

2. 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.

3. Evaluating the semantic relationship 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.

4. Based on the previous observations, providing a decentralized approach for computing the global document ranking in decentralized architecture for Web and P2P search [23] and report 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.

5. 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 search servers agree on a universal *SiteRank* in the document rank computation which allows to exclude spamming sites more easily.

In the next section we will provide arguments that the study of *SiteGraph*s could result in many benefits for Web information retrieval. Afterwards we introduce our model and the algorithm to compute *SiteRank* in Section 2. We did several sets of experiments to evaluate the significance of this idea. In Section 3, 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 that we made from these experiments, we

[2] Depending on the context, we use the same term *SiteRank* for both the algorithm and the rank value of a Web site.

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. To that end we summarize results from a companion paper [1] where we laid out the formal foundations for the distributed computation of rankings and we elaborated an example in that framework, with focus on the influence of *SiteRank* on the computation of document rankings. Finally, after a short review on related work, we draw some conclusion from our work and look into future research possibilities in Section 6.

2 A New Web Graph Leads to A New Rank

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

2.1 Random Walks in SiteGraph

We use the notion of *SiteLink* to designate hyperlinks among Web sites and *PageLink* for those among Web documents. Studies show that among the tens of billions PageLinks in the Internet, roughly 76% link to pages on the same Web host [4]. The estimated average distance between hosts has been found to be less than 6 (at most about 5.27).

Our algorithm is based on the random walk model in the *SiteGraph* which is similar to that of PageRank for *DocGraph*. Intuitively, a random walk models a simple process of randomly navigating in the Web (sometimes also called a "Drunkard's walk"). 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.

However, we emphasize the differences between our model for *SiteGraph* and the random walk model used in PageRank and many other related work on the *DocGraph*:

1. While PageRank considers users' navigating from one Web page to another, we are considering the access patterns of users at the granularity of Web sites. In the PageRank model, a user is assumed to be able to navigate to any other document, which is not linked with the current document, with constant probability. This is in reality impossible. A user could not even know the other 3 to 5 billion Web pages, thus assigning uniformly distributed weight to unknown pages is not a good approximation of the real situation on the Web. On the other hand, the number of Web sites is much smaller. A uniform distribution of weight to the much smaller number of Web sites appears to be more reasonable.

2. There are many implicit but empirically fully valid and frequent practices of navigating among Web pages which can not be properly reflected by the static *DocGraph*. One example is navigation from some internal page to the homepage of the Web site by removing the local navigation path of the document from the URL. This is possible even if there is no direct

link from the internal page to the homepage. As a result, if a Web site has on average 100 [3] documents, a modified *DocGraph* with 100 such implicit links added would be a better model of the the average Web site. Such de facto practices make PageRank's random walk model a skewed model of the structure of the real Web.

The situation is different in the random walk for Web sites. A user cannot simply take shortcut transformations since modifying URLs does not induce a possibility to navigate to another Web site. Thus the Web site is a more adequate conceptual and organizational unit for the Web, and the random walk model is quite suitable for the resulting *SiteGraph*.

3. 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 Web pages[4]. As a direct result the computed ranking results by algorithms like PageRank or HITS [15] 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.

2.2 The Algorithm

The representation of the *SiteGraph* by a matrix M_S is very similar to that for a *DocGraph*. The only difference is that every element in the *SiteGraph* matrix represents the number of *SiteLink*s instead of *PageLink*s. Self-referential links are also possible and counted and are represented as diagonal elements of the matrix.

Such a matrix may not have a non-trivial Eigenvector, which is the basic property of the matrix used by the PageRank algorithm. This may occur, for example, if there are isolated sets of pages that no other pages point to or there are dangling pages which point to no other page. One of the reasons that there exist such pages is that other pages pointing to or pointed by such pages have not been crawled. Different means have been devised to correct this problem. In the original PageRank paper [20], dangling pages are first of all removed. In other approaches this problem is addressed by assuming that users would often backtrack their navigation path, such that virtual links from the dangling pages to the pages pointing to them could be added to the link structure [21].

However, both methods do not address the problem of isolated sets of pages, which might occur when starting to crawl from a not fully connected set of seed sites. Moreover, we intentionally want to keep dangling pages without removing them in a preprocessing step. The reason is that although we have crawled a considerable number (more than 2 million) of our campus Web pages, there exist some sites that have only a single page having been crawled. If we remove such single pages, also the site will be removed from the derived *SiteGraph*. To obtain a *SiteGraph* that is as large as possible we keep those pages on purpose.

[3] This is roughly the average number of pages that a Web site in 2001 has. Per the work of Bharat et al. [4], the June 2001 Web data set has 1,292 millions of pages and 12.8 millions of nodes.

[4] Andrei Broder gave an interesting example in his keynote speech at the SIGIR'03 conference.

As in the Page Rank algorithm we apply the technique of introducing a decay factor to the original *SiteGraph*:

$$M_S = p \times M_S + \frac{1-p}{m_S} \times I$$

where m_S is the number of Web sites in the crawled graph, the decay factor p is 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.

It is easy to show that the resulting matrix M_S is both aperiodic and a irreducible stochastic transition matrix. As the random teleportation introduces non-zero weight self loops it is clear that the matrix is aperiodic. It is irreducible iff M_S is strongly connected. This is also guaranteed by the operation of applying a decay factor to the original matrix. Thus according to the Ergodic Theorem for Markov chains, the Markov chain defined by M_S has a unique stationary probability distribution. We can then compute the Web site ranking of this *SiteGraph* based on the matrix M_S.

Given the matrix representation M_S of the *SiteGraph* of the EPFL campus Web, and a start rank vector $\boldsymbol{r}^{(0)}$ of the Web sites in this *SiteGraph*, we apply the standard Power Method for computing the principal eigenvector to obtain the ranking for the Web sites:

function **PowerMethod**$(M_S, \boldsymbol{r}^{(0)})$ {
  $A = {M_S}^T$;
  repeat
    $\boldsymbol{r}^{(l+1)} = A\boldsymbol{r}^{(l)}$;
    $\delta = \|\boldsymbol{r}^{(l+1)} - \boldsymbol{r}^{(l)}\|_1$;
  until $\delta < \epsilon$;
  return $\boldsymbol{r}^{(l+1)}$;
}

3 The Significance of SiteRank

3.1 Data Set

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 two 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 *SiteLink*s 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 $M_S(i, i)$.

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. Our data set shows typical power-law properties: the fraction of pages having PageRank r is proportional to

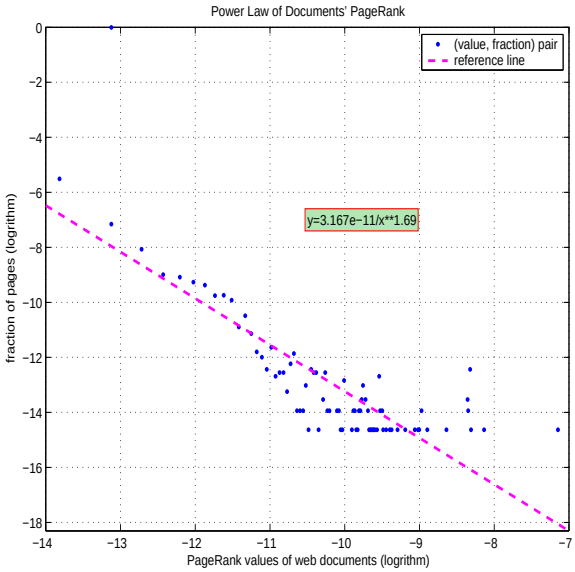

Figure 1: Log-log plot of the PageRank of the EPFL domain (.epfl.ch). A very typical power-law with exponent of about 1.69.

$1/r^{1.69}$. This result is strikingly similar to that reported in a study on the Web structure [21]. Though the exponent here is a bit lower than the value found there which is around 2.1. Two reasons might account for this difference: the difference in the nature of the different Web data sets we use; the incomplete crawling of our campus Web.

3.2 Power-Law Comes Back in SiteRank

In Figure 2, 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. Both axes are displayed at a logarithmic scale.

One of the interesting results of our work is that we found the *SiteRank* distribution also follows the power-law quite well: the fraction of Web sites having *SiteRank* r is proportional to $1/r^{0.95}$. Thus *SiteRank* becomes yet another property of the Web graph that abides by the powerful power-law.

However, the exponent of the *SiteRank* distribution is lower than that of the PageRank distribution. It would be interesting to analyze models for generating and growing the Web to obtain this empirically determined distributions and to capture the distinctive nature of the *DocGraph* and the *SiteGraph*. However this is out of the focus of this paper and remains part of our future work.

3.3 Closer Look in Terms of Semantic Quality

Next we show the top ranked Web sites according to the computed *SiteRank* values in Table 1. We omit the protocol prefix "http://" and the domain suffix ".epfl.ch" of all resulting URLs.

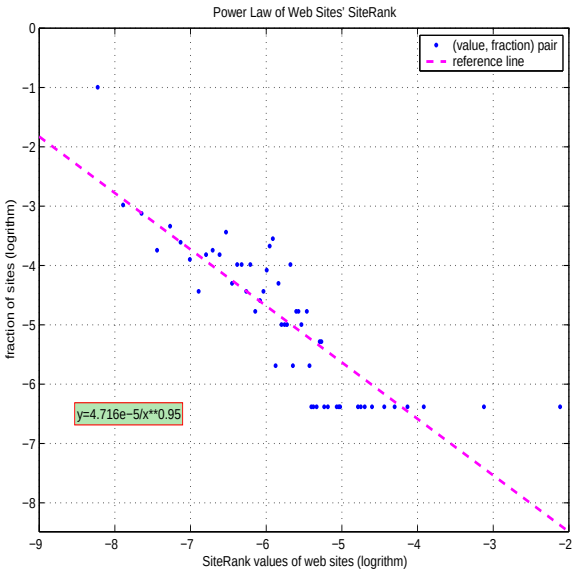

Figure 2: Log-log plot of the SiteRank of the EPFL domain (.epfl.ch). Also a typical power-law with exponent of about 0.95.

Order	SiteRank	Description
1	www	campus portal
2	spi	press and information
3	ic	school of comp. & comm. sciences
4	sti	school of engineering sci. & tech.
5	enac	school of environment, nature, architecture & construction
6	sb	school of basic sciences
7	cdh	school of humanities
8	sv	school of life
9	research	as the name indicates
10	plan	campus map
11	mediatheque	media library
12	daawww	administrative site
13	gaspar2	unified interface for identification and authentication
14	sicwww	computing center
15	vpf	vice president for education

Table 1: EPFL SiteRank

Some sites, for example daawww.epfl.ch, which is a site containing administrative information on the school, have not a standard index page and can thus not be directly accessed at the top site level. However, there are a large number of Web documents contained in its subdirectories, for example http://daawww.epfl.ch/daa, and thus its resulting *SiteRank* is not low at all.

One of our central ideas when using *SiteRank* is the possibility to use different rankings obtained from different contexts, sources and ranking methods. Each of the rankings represents a different interpretation of the concept of importance on the Web. To our knowledge exploiting this possibility for producing different, potentially personalized rankings has not been studied in the literature.

In order to explore the potential of such an approach we produced an alternative ranking of the schools' Web sites in Table 2 by using Google. We searched on Google using the keyword "EPFL". As "automated queries" are not allowed to be submitted to most of the commercial search engines[5] [9] [11], we manually submitted the keyword "EPFL" to Google, and gathered the top search results[6]. We then sieved the entries by only keeping the top 15 URLs site homepages found on the EPFL site. Unsurprisingly, we actually find that almost all the returned top entries are site homepages except for one, the Advanced Instant NT Password Cracker made by the lab of security and cryptography [17] at http://lasecpc13.epfl.ch/ntcrack/, probably because of the popularity of this tool. We compare the resulting list with our *SiteRank* based on the local campus Web graph.

A comparison of these two tables reveals several interesting facts:

1. In the Google result, two homepages (ltswww, and liawww) are only listed by their URLs and no text summary is attached, which implies that their page content has actually not been indexed by the search system[7]. Among them, only liawww set up a robots.txt file which merely disallows the crawling of cgi-bin files and a person's homepage. Such inaccuracy of information is a phenomenon that we have frequently observed for huge centralized search engines.

2. Moreover, three highly ranked homepages dmawww, dawww, dmtwww in the Google result were actually no longer used as the homepages of their units from late 2002 as the result of an academic reorganization in the school. Users inside the school should no longer use them to gather recent or official information. They are likely to be still ranked highly by Google because they are still referred to by many other Web pages of Web sites outside or even inside of the campus. This is another common phenomenon, that is hard for centralized search engines to address: information is often diffused or propagated at a very slow pace.

3. In the result based on the local Web graph, the entries appear to be more reasonable as most of them are notably the homepages of different

[5]In Google's terms, "sending automated queries" includes, among other things: using any software which sends queries to Google to determine how a Web site or Web page "ranks" on Google for various queries; "meta-searching" Google; and performing "offline" searches on Google. Also the company asks people "Please do not write to Google to request permission to 'meta-search' Google for a research project, as such requests will not be granted." [9]

[6]Search done on November 4, 2003, Tue..

[7]You get to know this because you do not find a "In Cache" link below or the cache is empty.

Order	Google	Description
1	www	campus portal
2	ligwww	virtual reality lab
3	dmawww	department of mathematics
4	icwww	school of comp. & comm. sciences
5	visiblehuman	visible human server
6	ltswww	signal processing institute
7	library	school central library
8	gnuwin	building gnu softwares for win
9	lslwww	logic system lab
10	dawww	section of architecture
11	liawww	artificial intelligence lab
12	dmtwww	department of microtechnique
13	lglwww	software engineering lab
14	asl	autonomous systems lab
15	plan	campus map

Table 2: Search "EPFL" on Google

organizational units. For internal users, this would be a perfect image of the site importance and corresponding page importance. But for external users, it is usually not. So obviously, a centralized search engine based on a single ranking scheme cannot serve the needs of two groups of users at the same time. And neither group's preference is trivial or can be neglected.

We have to keep in mind that the comparison above does not imply that one of the two rankings is better than the other for the following reasons:

1. The two rankings are based on different data sets. The Google ranking should be based on a recent crawl (presumably in October of 2003) of the EPFL campus Web. One of the limitations of crawling the Internet is that restricted Web sites are only accessible through the Intranet. On the other hand our *SiteRank* does not include the effects of external pages pointing to EPFL pages and it is not fully complete yet. How much this affects the ranking is an issue that requires further investigation.

2. A comparison of search results from Yahoo, AllTheWeb [10], and WiseNut [18] using the same query keyword "EPFL" would reveal that the results from different search engines are quite different, both with respect to results returned and ranking. This is expected as different search engines employ different ranking strategies.

These observations support our hypothesis that using *SiteRank*s might be an interesting approach, not only to realize a decentralized search framework, but also for the sake providing more meaningful rankings for specific usage contexts, by combining different site rankings and different local rankings.

3.4 PageRank in Relation to SiteRank

Our third set of experiments was conducted for elucidating the relationship between a document's PageRank and the *SiteRank* of the Web site the document

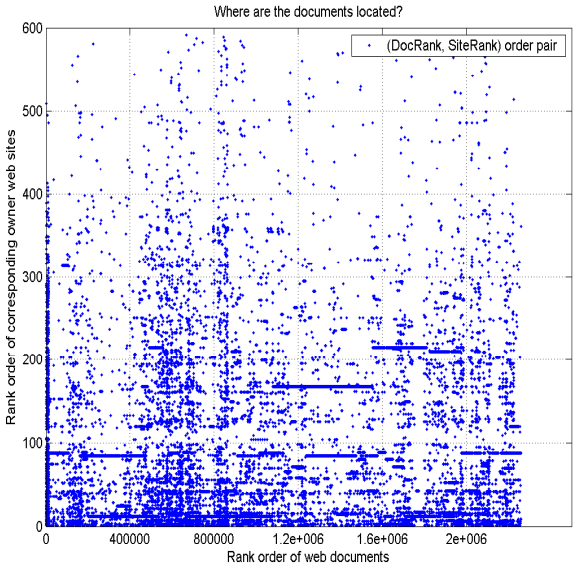

Figure 3: PageRank vs. SiteRank

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 the following, when we say a document is important in a general sense, we refer to its PageRank; when we say a Web site is important in a general sense, we refer to *SiteRank*.

In Figure 3, we display all (PageRank, *SiteRank*) order pairs. For example, if a document's PageRank order number is 1 which means it is ranked as the top 1 page, and the *SiteRank* order number of its owner site is also 1 which means its owner site is ranked as the top 1 site, the point (1, 1) is drawn in the diagram.

One can derive from this figure the following observations:

1. Highly ranked pages are distributed both at highly ranked Web sites and lowly ranked Web sites. This is not surprising however since even a highly ranked Web site like any well-known news agency Web portal could contain many less important pages which would not have very high PageRank values.

2. For several Web sites there are some prominent stripes in this diagram. They are in fact the documents having the same PageRank values. Since each page occupies one position in the rank list, they together form a stripe. They tend to appear in Web sites with higher *SiteRank*s as these sites tend to have a larger number of documents.

As Internet users tend to only notice the existence of highly ranked results from search engines, we specifically look at the distribution of top ranked documents on the Web sites. We show the diagram of the 1000 documents with the highest PageRank's[8] together with their corresponding Web sites in Figure 4.

[8]Please note that actual search engines return results ranked not only by PageRank, but

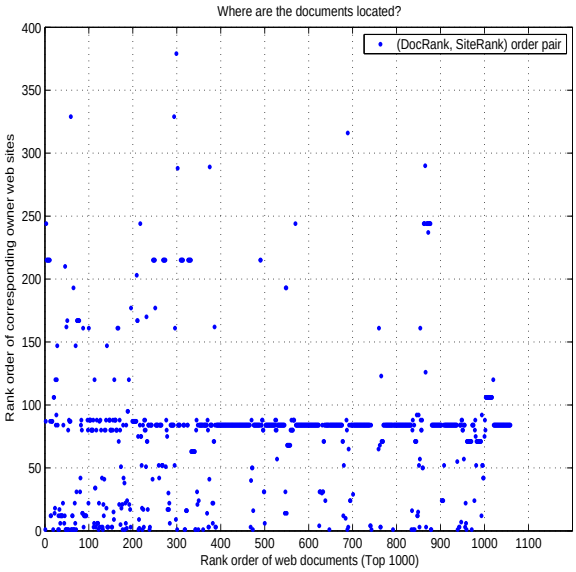

Figure 4: PageRank vs. SiteRank for the Top 1000 documents

There, we see 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 acutally a correlation between a page's rank value and the *SiteRank* of its owner. The stripe in this figure is the result of the same effect described for last figure.

It would be helpful to apply some standard method of computing the correlation between two vectors, such as Pearson's correlation coefficient, to study the correlation between the documents' PageRank values and the corresponding *SiteRank* values of their owner sites. However, such a method cannot be simply applied here as it assumes that both variables are approximately normally distributed and their joint distribution is bivariate normal. This assumption does not hold in our problem setting since we have shown that both the PageRank and SiteRank follow a power-law distribution. Furthermore, Spearman's correlation is not suitable here either since the sizes of the 2 ranking vectors are different. Even though, we can make non-quantitative observations on the relationship in between which is exhibited in Figure 5.

In Figure 5, the y dimension represents the SiteRank order number of a Web site, and the x dimension the lowest PageRank order number among all the Web documents on this particular Web site. The trend is very clear: the lower the SiteRank order number of a Web site (thus it's a highly ranked site), the lower the lowest PageRank order number of the documents (thus a highly ranked document) on the same site.

Based on the experimental results and observations above, we believe our assumptions below are very reasonable:

1. Many pages of an important Web site are also important.

---

usually a composite rank value also related to keyword matching, etc..

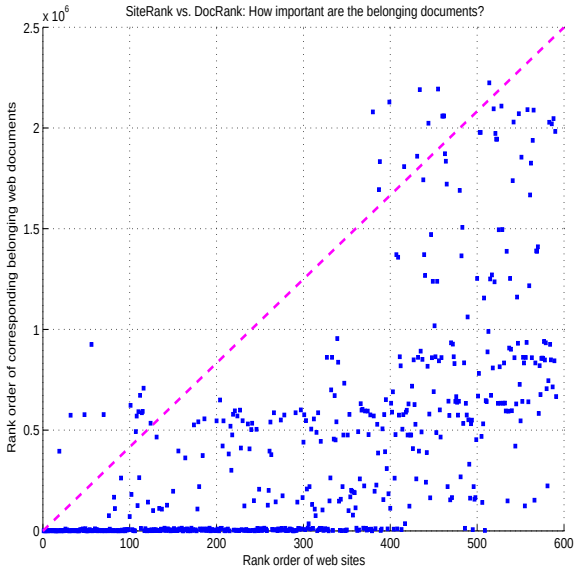

Figure 5: SiteRank vs. PageRank: minimal PageRank order numbers among the belonging documents of a Web site.

2. If a Web site has many important pages, it's highly probable that it is an important Web site.

Please note that these two statements are not tautological. If these statements hold true in a general sense or even if it is 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.

4 SiteRank for Decentralized Rank Computation

We want to distribute the task of computing Web document ranking to a set of distributed peers each of which crawls and stores a smaller 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*. For the sake of clarity, we call each of the participating servers in such a decentralized search system a *member server*.

4.1 SiteRank in a Big Picture

To fulfill our aim, we propose a decentralized architecture for search systems [23]. We decompose our computation of the global ranking for Web documents into three steps: the computation of *SiteRank*, the computation of local ranking of Web documents, and the application of the ranking algebra [1] to combine both rankings to produce the final global ranking.

We introduced a formal, algebraic framework for rank composition in [1], and demonstrated by a case study the utility of this approach to produce different forms of rankings. A key operator in this algebra is the *folding operator* which allows to combine site and local rankings into global rankings. We give a brief overview of this novel operator here, and refer the readers to [1] for the complete specification of the ranking algebra.

In order to operate with rankings at different abstraction levels, we introduce partitions of all Web documents. For example, P_S is a partition of all documents at the site level, whereas P_0 is a partition at the individual document level. A ranking can the be defined over a subset of the elements of a partition. In order to compare rankings at different abstraction levels we introduce a *covering relation* to relate partitions at different levels of granularity, e.g. $P_S \gg P_0$, which expresses that each element of P_0 is a subset of some element of P_S. A *covering vector* $R_{P_0}^{P_S}$ provides a partial relation among the elements of the coarser partition P_S, and (partial) rankings at the finer partition P_0. Thus a covering vector represents the relationship used to combine site and document rankings. The prototypical case is where the document ranking associated with one element $s \in P_S$ ranks exactly all the documents contained in s so $R_{P_0}^{P_S}(s)(d)$ is then the ranking for the document set d, which contains a single document, using the ranking associated with sites. However other more general cases are not excluded. Given a covering vector and a site ranking R_{P_S} the folding operation is then defined by:

$$R^*_{P_0}(d) = \sum_{d \in P_0,\ s \in P_S} R_{P_S}(s) * R_{P_0}^{P_S}(s)(d)$$

such that $d \subset s$, $R_{P_0}^{P_S}(s)(d)$ and $R_{P_0}(s)$ are defined.

The ranking R^* needs to be normalized to obtain the result of folding $\mathcal{F}^{P_S \gg P_0}(R_{P_0}^{P_S}, R_{P_0})$. More details on the complete algebra and its use are found in [1]. In this paper we focus on how to make use of *SiteRank*s in such a setting.

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 [4]. 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, a typical *SiteGraph* matrix M_S requires about $(591/2259102)^2 = 6.8 \times 10^{-6}\%$

memory or disk space as compared with the *DocGraph*. Moreover, we can use a 2-byte integer to represent every site, whereas we have to use at least a 4-byte integer or even a 8-byte integer for documents, this leads to another 50% or 25% saving of memory size or disk space. 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.

4.2 Case Study

We study the behavior of one specific member server in our decentralized search architecture. It crawls sicwww.epfl.ch, the home of the computing center (280 documents) and sunwww.epfl.ch, the support site for SUN machines (21685 documents) of our campus Web. We will compute the document ranking for these two sites to be studied. In order to bring in effects of external links, we also include the campus portal www.epfl.ch (2838 documents) as a reference. In all the computations we use the PageRank algorithm with different sets of documents to be computed. We compute the following different document rankings for the documents belonging to these two Web sites:

1. Ranking of all documents of the campus, which is used as a baseline. We compute the PageRank of all documents in the campus Web and project the result to the set of documents belonging to one of our selected sites to obtain the campus ranking. All the following ranking results will be compared with this baseline.

2. Subset ranking which is obtained by computing PageRank on the documents belonging to one of the above three Web sites including the reference site and then projecting the result to the documents of the two selected sites.

3. Tinyset ranking which is the result obtained by only selecting the documents belonging to the two selected sites and applying the PageRank algorithm to them.

4. Random ranking where each document is assigned a random rank value to keep the sum of all rank values 1.

5. Composite ranking: Each one of the two selected Web sites first computes its own local PageRank based only on its own local document set. This local ranking is then combined with the rank weights endorsed by external links originating from the other two sites. External Web links are weighted by the *SiteRank*s of their source Web site.

The distances between the last four ranking vectors and the baseline vector are shown in Figure 6. The distance measure we use is a weighted Spearman's Footrule:

$$F(R_0, R_1) = \sum_{i=1}^{n} w_0(i) w_1(i) |R_0(i) - R_1(i)|$$

In the formula, $R_j, j = 0, 1$ are the two ranking vectors to be compared. $R_j(i)$ is the rank order of document i. Since users mostly care about top listed documents we assign 90% of the weight to the T top-listed documents for $T < n$,

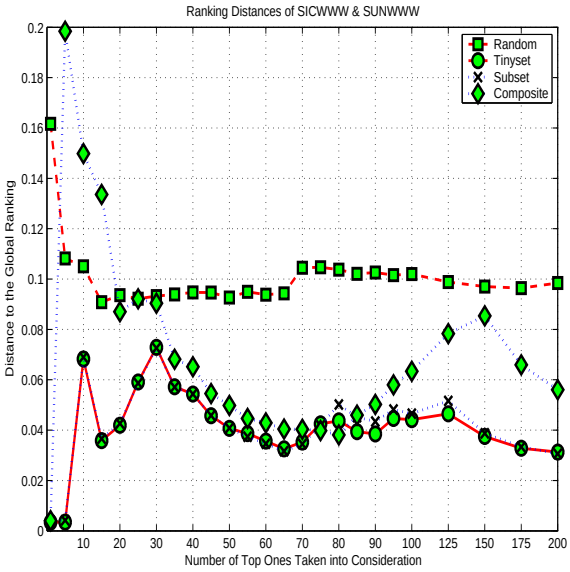

Figure 6: Ranking Distances of SICWWW & SUNWWW

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 $1 \leq i \leq n$.

One can observe that as soon as a larger number of top-ranked documents is taken into account, the composite ranking approximates the rankings computed on the selected subsets. This is an interesting result, since the composite ranking is performed in a distributed manner, making use of *SiteRank*, computing separate rankings for each of the three subdomains involved, whereas the *subset* and *tinyset* rankings can be considered as corresponding to a PageRank ranking based on the union of the selected subdomains, taking into account cross-site links. This shows that by composition one can obtain at least as good results in a distributed manner as with global ranking using the same information. Also interestingly, the result of the *composite* ranking appears to be *worse* if only the very few first result documents are considered. However, by inspecting the results in more detail, one observes that these results are actually qualitatively better. In the composite ranking more meaningful URLs, such as homepages of organizational units are included, while the global rankings return in particular pages, such as JDK documentation as the top ones, simply because they heavily cross-link to each other.

5 Related Work

Many previous studies on the structure and the distribution properties of the Web as a graph have been performed [16, 5]. Different snapshots of the Web have been investigated to find that the in-degree distribution can be described by the function $c_i/k^{2.1}$ where k is the in-degree number and c_i is a normalization constant so that the fractions sum to 1. In parallel, the out-degree distribution can be approximated by the function $c_o/k^{2.7}$ where k is the out-degree num-

ber. Such distributions are known as the power-law of the Web [7]. Recent research [21] found that not only the page in-degree, page out-degree, but also the PageRank values follow this power law as well. 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.

Obviously the work we presented here is tightly related to the rank computation for Web documents. Ever since the birth of PageRank, the method and algorithm for computing the ranking or degree of importance for Web pages has become an attractive and important research topic. Many recent results exist. Abiteboul et al. devised a way of performing on-line page importance computation [2]. The application of the PageRank algorithm to Peer-to-Peer systems was also proposed [22] by Sanharalingam et al.. However, in their approach they just replace the cost-intensive iterated matrix multiplication, by a slower on-line propagation and integration of the rank values of individual peers' collection of documents. This approach would suffer from sensitivity with respect to the choice of initial rank values, slow convergence because of network connection delays, and the huge number (tens or even over a hundred million) of messages that have to be generated to forward the rank information. We looked at an approach of how to perform PageRank computation incrementally for ephemeral documents those who come into being in between two consecutive crawls of the Web [24].

In addition to the question of how to compute the ranking in different distributed architectures, many methods for speeding up the original PageRank algorithm have been proposed [8, 13, 12]. For example, BlockRank [14] makes use of the block structure of Web data, but it is still a centralized algorithm to compute the PageRank, which is radically different from our decentralized approach. In contrast, the work we present in this paper is based the *SiteRank* rank for Web sites and as such takes advantage of preexisting structures in the Web graph.

Hyperlinks among Web sites were also studied [4] to reveal relationships among them and the properties of the host link structure, e.g., Zipfian distributions of weighted host in-degree and out-degree at this coarser host granularity, the average degree of separation, linkage of hosts in different top-level (including country) domains, etc.. Yet to the best of our knowledge, the study of the Web *SiteGraph* has not been previously used in the computation of Web document rankings for search engines. Our work shows that the *SiteRank* distribution also follows the power-law and then explores the possibility of how to apply it in a decentralized framework [23] to compute the global Web document ranking.

6 Conclusion and Future Work

In this paper, we first identify different abstractions for the Web graph depending on the granularity. We introduce *DocGraph* and *SiteGraph* at the level of Web documents and the level of Web sites respectively. After that we present the algorithm to compute the *SiteRank* for a *SiteGraph* which is a variant of the PageRank algorithm. We then study the graph properties of our crawled campus Web. We find that the PageRank distribution follows the power-law which is similar to recent results reported by other researchers. More interestingly we find that also the *SiteRank* roughly follows the power-law as well, with

a different exponent though. We then reveal some useful correlation between the PageRank and *SiteRank*. Based on these results and observations, 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.

Previous work [6] on the self-similar behavior of the Web (properties of the global Web are reflected in sub-domains and smaller subgraph snapshots) provides some basis for our experiments. Thus we think that the case study related to a campus-scale Web is to a certain degree representative to demonstrate the relationship between PageRank and *SiteRank*s of the Web sites that the documents belong to. However, an obvious future work is to test the idea with a huge scale data set, for example, the *SiteGraph* of the Swiss Web, or even the whole Web. With a sufficiently large snapshot of a Web subgraph, we will also no longer have the problem of having a limited number of Web sites, so we can remove the dangling pages as what the original PageRank algorithm did in a preprocessing step to see if the results will be of much difference.

Our work shows that the *SiteRank* of a Web subgraph also exhibits the property of the power-law though with a lower power parameter of about 0.95. As this is the first report on a *SiteRank* power-law parameter, we look forward to refine the result in future experiments with new data sets. Another related issue is the Web growth model. Previous research on such models focused on how a new Web page is added to the existing Web graph, however this is not the real situation in the Web. On the contrary, Web administrators have to take account of the set up of subdomains and corresponding Web sites before pages are created and added. Thus, we may study Web growing models at this higher granularity of Web sites. In addition, formal methods will be required to study the correlation between the PageRank and *SiteRank* vectors.

Further research related to existing major search engines would also be of interest. The subsets of documents covered by different Web search engines vary widely. It was reported that only 1.4% of the total coverage Web indexed by all four major search engines (HotBot, AltaVista, Excite, and Infoseek) in late 1997 [3]. Later research in 2002 [19] shows that in 141 hits of four small searches run on ten different engines, only 30 were found by two. Thus even using the same algorithm PageRank, the resulting ranking of the documents in respective search engines would be different. However, it would be a reasonable assumption that the sets of Web sites crawled by the search engines would exhibit a substantially higher overlap, since the number of sites is two magnitudes lower than that of documents. Thus *SiteRank*s computed on several existing commercial search engines may not be much different from each other. It would be interesting to compare such ranking characteristics among search engines.

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 the twelfth international conference on World Wide Web*, pages 280–290, Budapest, Hungary, 2003. ACM Press.

[3] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proc. of the 7th World-Wide Web Conference (WWW7)*, 1998.

[4] 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.

[5] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320, Amsterdam, The Netherlands, 2000. North-Holland Publishing Co.

[6] Stephen Dill, S. Ravi Kumar, Kevin S. McCurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. Self-similarity in the web. In *The VLDB Journal*, pages 69–78, 2001.

[7] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[8] Taher Haveliwala. Efficient computation of pageRank. Technical Report 1999-31, Stanford University, September 1999.

[9] Google Inc. Terms of service. http://www.google.com/terms_of_service.html, visited on 13th of Oct. 2003.

[10] Overture Services Inc. All the web search engine. http://www.alltheweb.com/.

[11] Overture Services Inc. Terms of use. http://www.alltheweb.com/info/about/terms_ of_use, visited on 13th of Oct. 2003.

[12] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[13] Sepandar Kamvar, Taher Haveliwala, and Gene Golub. Adaptive methods for the computation of pagerank. Technical report, 2003.

[14] Sepandar Kamvar, Taher Haveliwala, Christopher Manning, and Gene Golub. Exploiting the block structure of the web for computing pagerank. Technical report, 2003.

[15] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[16] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, 1627:1–17, 1999.

[17] EPFL LASEC. The security and cryptography laboratory (lasec). http://lasecwww.epfl.ch/.

[18] LookSmart Ltd. Wisenut search engine. http://www.wisenut.com/.

[19] Greg R. Notess. Search engines statistics: Database overlap. http://www.searchengineshowdown.com/stats/overlap.shtml, visited on 10th of Nov. 2003.

[20] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, January 1998.

[21] Gopal Pandurangan, Prabhakara Raghavan, and Eli Upfal. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.

[22] Karthikeyan Sankaralingam, Simha Sethumadhavan, and James C. Browne. Distributed pagerank for p2p systems. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC'03)*, pages 58–290, Seattle, Washington, USA, June 2003. IEEE Computer Society. June 22-24, 2003.

[23] Jie Wu. Towards a decentralized search architecture for the web and p2p systems. In *Proceedings of the Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2003), The fourteenth conference on Hypertext and Hypermedia, HyperText 2003*, Nottingham, U.K., August 2003.

[24] Jie Wu and Karl Aberer. Incrementally ranking ephemeral web documents in search engines. In *Mathematical/Formal Methods in IR, Workshop in SIGIR'03*, Toronto, Canada, August 2003.