A Framework for Decentralized Ranking in Web Information Retrieval

Karl Aberer and Jie Wu

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

Abstract. Search engines are among the most important applications or services on the web. Most existing successful search engines use global ranking algorithms to generate the ranking of documents crawled in their databases. However, global ranking of documents has two potential problems: high computation cost and potentially poor rankings. Both of the problems are related to the centralized computation paradigm. We propose to decentralize the task of ranking. This requires two things: a decentralized architecture and a logical framework for ranking computation. In the paper we introduce a ranking algebra providing such a formal framework. Through partitioning and combining rankings, we manage to compute document rankings of large-scale web data sets in a localized fashion. We provide initial results, demonstrating that the use of such an approach can ameliorate the above-mentioned problems. The approach presents a step towards P2P Web search engines.[1]

Keywords: search engines, information retrieval, P2P systems, link analysis

1 Introduction

A number of papers in recent years have studied the approach of hyperlink structure analysis to determine the hub and authority values or reputation of web documents. One algorithm proposed, PageRank [2], is the kernel of the very successful search engine Google [1]. The main problem of this sort of solutions is that they still are based on a centralized framework. The computation cost is thus prohibitively high since they have to deal with the complete document collection of the whole web. Even worse, the computation result can never reflect precisely the *real*[2] ranking relationship since the snapshot in search engines' databases can never be complete.

[1] 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.

[2] That is, the ideal ranking of all existing web documents on the Internet.

X. Zhou, Y. Zhang, and M.E. Orlowska (Eds.): APWeb 2003, LNCS 2642, pp. 213–226, 2003.

We give a more detailed analysis that strongly suggests that increased use of local rankings should be made as compared to global rankings. Computing local rankings not only allows to partition the problem of determining a global ranking and to derive this ranking from fresher information, but also allows to peruse information that is only locally available for the ranking computation. Examples of such information are the hidden Web and usage profiles. In this paper, we also take up the idea of sharing resources at the level of both computing and knowledge in P2P systems, and explore the possibility to determine document rankings for the use in Web search.

We first point out some of the potential that such an approach bears, such as better scalable architectures and improved usage of distributed knowledge. The key in making such an approach work lies in the ability to compose global rankings from local rankings. We develop an *algebra* to concisely specify ranking composition processes. An important contribution of the paper is the identification of the essential operators for ranking compositions that we have derived from experiences gained by composing rankings in an ad-hoc manner in our experimental environment. We then present some initial results on ranking compositions, that demonstrate their potential use: on the one hand in decomposing global ranking computation and still being able to retain the original ranking characteristics, on the other hand in composing new types of rankings depending on the desired ranking context.

The paper lays ground for many possible future developments. Most notably, these concern system architectures that allow to implement ranking composition processes efficiently and the use of the algebraic framework for potential optimizations of ranking compositions by algebraic rewriting.

2 Limitations of the Global Ranking Approach

In this section we give an overview of a number of inherent limitations of the global ranking approach, as for example used by Google. Some of those are more or less well-known facts, whereas others will be more surprising. The limitations fall into three categories:

1. Practical problems related to scalability
2. Semantic problems related to the exclusive use of global context
3. Semantic problems related to the instability of ranking algorithms

We will illustrate the problems at the example of Google, in particular the use of the PageRank algorithm, but similar problems can be expected for other ranking approaches. PageRank uses iterations of matrix multiplication on the global web adjacency matrix to compute a global ranking of the whole web graph, and returns the search results in an order that is decided by a combined factor of PageRank and the ranking of the traditional keyword-based search.

2.1 Practical Problems of Global Algorithms

We call the algorithms based-on global information "global algorithms". The practical problems of global ranking schemes are mainly related to scalability.

Dynamicity of Web content: According to the recent research result, the Web consists of approximately 2.5 billion documents in 2000, with a rate of growth of 7.3 million pages per day [7]. This web growth rate continuously imposes high pressure on existing search engines. Repetitive computation is required even if only a small part of the global web is changed. The reason is that, the global link adjacency matrix is required to compute the final PageRank.

Latency: Most search engines update on a roughly monthly basis [4]. Since the time needed to retrieve all the existing and newer Web increases, it also takes longer time to integrate it into the database, thus longer for a page to be exposed on search engines. A simpler way to demonstrate the impossibility of catching up with the growing speed of the Web for the current centralized search systems is that, in December 2001, Google announced that it was spidering 3 million pages each day where freshness had been determined to be crucial [10]. So the Web pages emerging per day are crawled by Google in about 2.5 days. As a consequence also the Web graph structure that is obtained will be always incomplete, and the global ranking computation thus less accurate.

High Computation Cost: The computation of PageRank is over the whole Web that has been grabbed into Google's servers. Therefore, the algorithm has to deal with the problem of multiplication of a huge matrices. Early in June 1 of year 2000, Google indexes 300 million pages in total, and Google's process entails 500 million variables and 2 million terms to index every month, resulting in about 1 terabyte of data to index. According to Sergey Brin, hypertext analysis of Google is computationally expensive [6]. A single matrix multiplication with 75 million URLs takes 5 hours. At that time, Google already used 4,000 PCs running Linux to provide its service.

2.2 Importance of Context for Ranking

Link-based ranking methods, such as Google's PageRank algorithm [2] and the hub and authority method by Kleinberg [3], have proven as a valuable approach in uncovering hidden, distributed knowledge in the Web. They are based on the implicit assumption that the existence of a link from a Web document to another document expresses that the referenced document bears some importance to the content of the referencing document and that frequently referenced documents are of a more general importance.

The rankings derived based on that observation are usually established in the context of a specific query, either in combination with other global ranking schemes as in Google or by post-processing query results as in Kleinberg's proposal. However, other forms of context may be considered, in particular the aspect of locality.

The first observation we make is that there exists a certain likelihood that a local link, i.e. a link that references a document within the same local domain,

typically a Web site, is likely to be semantically more "precise" since the author of the link is likely to be better informed about the semantics and particular importance of the local documents than an external author.

The second observation we make is that documents that are globally considered as important, also locally will have greater importance. This second observation suggests that it might be plausible to identify documents of global importance based on there local rankings only.

The third observation we make is that each Website establishes a specific semantic context. Depending now on the context we might specifically take advantage of the semantics implicit in certain Websites in order to obtain rankings that are tuned towards certain interest profiles.

All of these three observations lead us to the conclusion that it might be worthwhile to consider from a semantic perspective instead of a single global ranking various combinations of local rankings for the following three different but not mutually exclusive purposes:

1. Obtaining more precise rankings by exploiting local knowledge;
2. Reconstructing global rankings from local rankings in order to distribute the ranking effort;
3. Using selected local rankings in order to tune the resulting ranking towards specific interest profiles.

2.3 Instability of Ranking Schemes

Little attention is generally paid to the question to which extent link-based ranking methods are sensitive to changes in context. We performed a number of experiments indicating that link-based ranking such as PageRank might have in fact some undesirable properties with respect to stability. We classified them into two problems.

Effects of Agglomerate Documents

Previous studies on the HITS algorithm [9] revealed that HITS is prone to the problem of mutual reinforcement: the hub-authority relationships between pages are mutually reinforced because people put some one-to-many or many-to-one links in web sites. This problem can be solved in a heuristic way by dividing the hub or authority weights in the computation by the in-degree or out-degree number. It seems that it has not been noticed in the literature that the same phenomenon also occurs for the PageRank algorithm. The heuristic solution used by HITS to circumvent the problem cannot be applied to PageRank, since the division by the out-degree number is already used in the PageRank algorithm.

We illustrate this phenomenon by a simple experiment. We applied the PageRank algorithm to the set of documents that can be found at the ETH Zuerich website (about 430.000 pages). Interestingly, among the top 20 documents of PageRank one finds a substantial number of pages from the Java documentation (13 out of 20), which surely are not the most relevant documents to characterize ETH Zuerich. The reason for those documents to being ranked that high is found in the strong cross-referencing the Java documentation exhibits.

Stability of Local Ranking

Computation of global rankings merges information that is drawn both from local links and remote links. An interesting question is on the influence local versus remote links can have on the outcome of the ranking computation.

We illustrate this point by another experiment we did with pages collected from the EPF Lausanne websites (domain ".epfl.ch"). We chose two subsets of pages from them, related to two different organizational units and included all pages referenced from these web sites which brought the total number of documents to 1075. We computed now local rankings for documents for both websites (*dscwww* and *icawww*) in two ways.

1. Computing a global ranking including all 1075 documents and then projecting the resulting global ranking to the pages from one website;
2. Computing a local ranking from the documents found on each respective web site only.

The result is somewhat surprising. For the smaller (*dscwww*) of the two websites, both the projected global and the local ranking coincide almost completely. For the larger (*icawww*) of the two the projected global and the local ranking are substantially different. Analysis shows that by relying solely on global rankings different aspects of ranking semantics, namely the local ranking (self-assessment) and the projected global ranking (assessments by others) are merged in a somewhat arbitrary manner. Therefore a separation of these concerns seems to be a promising approach in order to reveal more precise information from the available link structure.

3 The Ranking Algebra

In the previous section we have argued that different rankings established in different contexts can be of interest. Thus we see rankings as first-class objects, that can be produced, exchanged and manipulated as any other data object. We introduce now a framework that defines what the type of rankings is, and how rankings are manipulated. We will use an algebraic framework for rankings, a ranking algebra, similarly as it is done for other types of data objects (such as using relational algebra for relations). The ranking algebra will allow to formally specify different methods of combining rankings, in particular, for aggregating global rankings from local rankings originating from different semantic contexts.

3.1 Definitions

First we have to define the domain of objects that are to be ranked. Since rankings can occur at different levels of granularity there will not be rankings of documents only, but more generally, rankings over subsets of documents. This leads to the following definition.

Definition 1: A partition of a document set D is a set P of disjoint, non-emtpy subsets of D where $P = \{p_1, \cdots, p_k\}$, $D = \bigcup_{i=1}^{k} p_i$. We denote $\mathcal{P}(D)$ or briefly $\mathcal{P}$ as the set of all possible partitions over the document set D. We call each of the disjoint subsets a zone.

We use $\mathbf{P_0}$ to denote the finest partition where each zone in it is a single web document. So rankings at the document levels are also expressed over elements of $\mathcal{P}$ which makes our ranking framework uniform independent of the granularity of ranking. We also use $\mathbf{P_S}$ to denote the partition according to web sites, assuming that there exists a unique way to partition the Web into sites (e.g. via DNS). Then each zone corresponds to the set of web documents belonging to an individual site.

In order to be able to compare and relate rankings at different levels of granularity we introduce now a partial order on partitions.

Definition 2: Given $\mathcal{P}(D)$, the relation *cover* over $\mathcal{P}(D)$ for $P_1, P_2 \in \mathcal{P}(D)$ is denoted as $P_1 \ll P_2$ and holds iff. $\forall p_1 \in P_1, \exists p_2 \in P_2, p_1 \subseteq p_2$.

We also say that P_1 is covered by P_2 or P_2 covers P_1. The relation $P_2 \gg P_1$ is defined analogously.

We will also need a possibility to directly relate the elements of two partitions to each other (and not only the whole partitions as with cover). Therefore we introduce the following operator.

Definition 3: For $P_1, P_2 \in \mathcal{P}, P_1 \gg P_2$ the mapping $\rho_{P_1 \gg P_2} : P_1 \to 2^{P_2}$ is defined for $p \in P_1$ and $q \in P_2$ as $q \in \rho_{P_1 \gg P_2}(p)$ iff. $q \subseteq p$.

This operator selects those elements of the finer partition that are covered by the selected element p of the coarser partition. For example, for $\mathbf{P_S} \gg \mathbf{P_0}$, given a web site $S \in \mathbf{P_S}$, the operator maps it to its set of web documents contained in this site: $\rho(S) \subseteq \mathbf{P_0}$.

The basis for computing rankings are links among documents or among sets of documents. Therefore we introduce next the notion of link matrix. Link matrices are always defined over partitions, even if we consider document links. Also we define link matrices only for sub-portions of the Web, and therefore introduce them as partial mappings. Note that it makes a difference whether a link between two entities is undefined or non-existent.

Definition 4: Given $P \in \mathcal{P}$ a link matrix $M_P \in \mathcal{M}_P$ is partial mapping $M_P : P \times P \to \{0, 1\}$. In particular if M_P is defined only for values in $P_1 \subset P$ then we write $M_P(P_1)$. We say then $M_P(P_1)$ is a link matrix over P_1.

A number of operations are required to manipulate link matrices before they are used for ranking computations. We introduce here only those mappings that we have identified as being relevant for our purposes. The list of operations can be clearly extended by other graph manipulation operators.

The most important operation is the projection of a link matrix to a subset of the zones that are to be ranked.

Definition 5: For $P \in \mathcal{P}(D)$, $P_1 \subseteq P$ and $M_P \in \mathcal{M}_P$, the node projection $\mathit{\Pi}_{P_1} : \mathcal{M}_P \to \mathcal{M}_P(P_1)$ satisfies $\mathit{\Pi}_{P_1}(M_P)(p, q), p, q \in P$ defined iff. $p, q \in P_1$ and M_P is defined for p, q.

We also need the ability to change the granularity at which a link matrix is specified. This is supported by the contraction operator.

Definition 6: For $P_1, P_2 \in \mathcal{P}(D)$ with $P_1 \gg P_2$ and link matrices $M_{P_1} \in \mathcal{M}_{P_1}$ and $M_{P_2} \in \mathcal{M}_{P_2}$ the contraction $\Delta^{P_1 \gg P_2}$: $\mathcal{M}_{P_2} \to \mathcal{M}_{P_1}$ is the mapping that maps M_{P_2} to M_{P_1} such that for $p', q' \in P_1$, $M_{P_1}(p', q')$ defined iff. $M_{P_2}(p, q)$ defined for all $p, q \in P_2$ with $p \subseteq p', q \subseteq q'$ and $M_{P_1}(p', q') = 1$ iff. $M_{P_1}(p', q')$ defined and exists $p, q \in P_2$ with $p \subseteq p', q \subseteq q', M_{P_2}(p, q) = 1$.

for $p, q \in P_2$ $M_{P_2}(p, q) = 1$ and defined iff. for $p', q' \in P_1$ with $p \subseteq p', q \subseteq q'$ $M_{P_1}(p', q') = 1$ and defined.

In certain cases it is necessary to directly manipulate the link graph in order to change the ranking context. This is supported by a link projection.

Definition 7: For $P \in \mathcal{P}(D)$, $P_1 \subseteq P$ and $M_P \in \mathcal{M}_P$ the link projection $\Lambda_{P_1} : \mathcal{M}_P \to \mathcal{M}_P$ satisfies for $p \in P - P_1, q \in P - P_1$ $\Lambda_{P_1}(M_P)(p, q) = 0$ iff. $M_P(p, q)$ defined and $\Lambda_{P_1}(M_P)(p, q) = M_P(p, q)$ for all other p, q.

Based on link matrices rankings are computed. The domain of rankings will again be partitions of the document set.

Definition 8: For $P \in \mathcal{P}(D)$ a ranking $R_P \in \mathcal{R}_P$ is a partial mapping $R_P : P \to [0, 1]$. When the ranking is defined for $P_1 \subseteq P$ only we also denote the ranking as $R_P(P_1)$.

Normally rankings will be normalized. This leads to the following definition:

Definition 9: A normalized ranking R_P satisfies $\sum_{p \in P} R_P(p) = 1$. Given a general ranking $R_P \in \mathcal{R}_P$ the operator $\mu : \mathcal{R}_P \to \mathcal{R}_P$ derives a normalized ranking by $\mu(R_P(p)) = \frac{R_P(p)}{\sum_{p \in P} R_P(p)}$.

The connection between rankings and link matrices is established by ranking algorithms. As these algorithms are specific, we do not define their precise workings.

Definition 10: A ranking algorithm is a mapping $R_P^{alg} : \mathcal{M}_P(P_1) \to \mathcal{R}_P(P_1)$

We will distinguish different ranking algorithms through different superscripts. In particular, we will use $R^{PageRank}$, the Page rank algorithm, and R^{Count}, the incoming links counting algorithm, in our later examples.

As for link matrices we also need to be able to project rankings to selected subsets of the Web.

Definition 11: For $P \in \mathcal{P}(D)$ and $R_P \in \mathcal{R}_P$ the projection $\Pi_{P_1} : \mathcal{R}_P \to \mathcal{R}_P(P_1)$ is given as $\Pi_{P_1}(R_P) = \mu(R_P^*)$ iff. $R_P^*(p) = R_P(p)$ with $p \in P_1$ and $R_P(p)$ defined.

In many cases different rankings will be combined in an ad-hoc manner driven by application requirements. We introduce weighted addition for that purpose.

Definition 12: Given rankings $R_P^i \in \mathcal{R}_P, i = 1, \ldots, n$ and a weight vector $w \in [0, 1]^n$ then the weighted addition $\Sigma_n : \mathcal{R}_P^n \times [0, 1]^n \to \mathcal{R}_P$ is given as $\Sigma_n(R_P^1, \ldots, R_P^n, w_1, \ldots, w_n) = \mu(R_P^*)$ iff. $R_P^*(p) = \sum_{i=1}^n w_i R_P^i(p)$ and $R_P^i(p)$ defined for $i = 1, \ldots, n$.

We will in particular look into methods for systematic composition of rankings. These are obtained by combining rankings that have been obtained at different levels of granularity. To that end we introduce the following concepts.

Definition 13: A covering vector of rankings for $\mathcal{R}_Q$ over $\mathcal{R}_P$ with $Q \gg P$ is a partial mapping $R_P^Q \in \mathcal{R}_P^Q$ with signature $R_P^Q : Q \to \mathcal{R}_P$.

This definition says that for each ranking value of a ranking at higher granularity there exists a ranking at the finer granularity. Next we introduce an operation for the systematic composition of rankings using covering vectors.

Definition 14: Given a covering vector $\mathcal{R}_P^Q$ with $Q \gg P$ the folding is the mapping $\mathcal{F}^{Q \gg P} : \mathcal{R}_P^Q \times \mathcal{R}_Q \rightarrow \mathcal{R}_P$ such that for $R_P^Q \in \mathcal{R}_P^Q, R_Q \in \mathcal{R}_Q$, $\mathcal{F}^{Q \gg P}(R_P^Q, R_Q) = \mu(R_P^*)$ iff. for $p \in P$,

$$R_P^*(p) = \sum_{q \in Q \text{ st. } R_P^Q(q) \text{ and } R_Q(q) \text{ defined}} (R_Q(q) * R_P^Q(q)(p)).$$

3.2 Computing Rankings from Different Contexts

In this section we give an illustration of how to apply the ranking algebra in order to produce different types of rankings by using different ranking contexts.

Suppose $P_S = \{s_1, \ldots, s_k\} \subset \mathbf{P_S}$ is a subset of all Web sites. If we determine $D_i = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_i)$ we see that $D_i \subset \mathbf{P_0}$ corresponds to the set of documents of the Web site s_i. We denote with $D_S = \cup_{i=1}^k D_i$ the set of all documents occuring in one of the selected Web sites. For ranking documents from the subset P_S of selected Web sites we propose now different schemes.

Global site ranking: The global site ranking is used to rank the selected Web sites using the complete Web graph. Since only inter-site links are used the number of links considered for computing the ranking is substantially reduced as compared to the global Web graph. In addition such rankings should only be recomputed at irregular intervals. The ranking algorithm to be used is PageRank. Global site rankings for subsets of Web sites could be provided by specialized ranking providers or Web aggregators. Formally we can specify this ranking as follows. Given the Web link matrix $M \in \mathcal{M}_{\mathbf{P_0}}$ and a selected subset of Web sites $P_S \subset \mathbf{P_S}$ the global site ranking of these Web sites is given as

$$R_{P_S}^{GS} = \Pi_{P_S}(R^{PageRank}(\Delta^{\mathbf{P_S} \gg \mathbf{P_0}}(M))) \in \mathcal{R}_{\mathbf{P_S}}(P_S)$$

Local site ranking: In contrast to the global site ranking we use here as context only the subgraph of the Web graph that concerns the selected Web sites. In this case we prefer to use the ranking algorithm R^{Count} since the number of inter Web site links may be more limited for this smaller link graph. Formally we can specify this ranking as follows. Given the Web link matrix $M \in \mathcal{M}_{\mathbf{P_0}}$ and a selected subset of websites $P_S \subset \mathbf{P_S}$ the local site ranking of these websites is

$$R_{P_S}^{LS} = R^{Count}(\Pi_{P_S}(\Delta^{\mathbf{P_S} \gg \mathbf{P_0}}(M))) \in \mathcal{R}_{\mathbf{P_S}}(P_S)$$

Note that we assume that R^{count} ranks only documents for which the link matrix is defined and thus we don't have to project the resulting ranking to the subset of Web sites taken into account.

Global ranking of documents of a Web site: This ranking is the projection of the global PageRank to the documents from a selected site. Formally we can specify this ranking as follows. Given the Web link matrix $M \in \mathcal{M}_{\mathbf{P_0}}$

and the Web site $s_i \in P_S$ with $D_i = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_i)$. Then the global ranking of documents of a Web site is

$$R_{D_i}^{global} = \Pi_{D_i}(R^{PageRank}(M))) = \Pi_{D_i}(R_{D_S}^{global}) \in \mathcal{R}_{\mathbf{P_0}}(D_i)$$

A more restricted form of global ranking is when we only include the documents from the set $D_S = \cup_{i=1}^{k} D_i$. This gives

$$R_{D_i}^{intermediate} = \Pi_{D_i}(R^{PageRank}(\Pi_{D_S}(M)))) \in \mathcal{R}_{\mathbf{P_0}}(D_i)$$

The global or intermediate ranking of documents of a set $D' = D_{i_1} \cup \cdots \cup D_{i_m}$ of more than one web sites can be obtained similarly by simply replacing D_i with D' in the projection operators.

Local internal ranking for documents: This corresponds to a ranking of the documents by the document owners, taking into account their local link structure only. The algorithm used is PageRank applied to the local link graph. Formally we can specify this ranking as follows. Given the Web link matrix $M \in \mathcal{M}_{\mathbf{P_0}}$ and the Web site $s_i \in P_S$ with $D_i = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_i)$, the local internal ranking is

$$R_{D_i}^{LI} = R^{PageRank}(\Pi_{D_i}(M)) \in \mathcal{R}_{\mathbf{P_0}}(D_i)$$

Note that we assume here that the PageRank algorithm does not rank documents for which the link matrix is undefined, and therefore the resulting ranking is only defined for the local web site documents.

Local external ranking for documents: This corresponds to a ranking of the documents by others. Here for each document we count the number of incoming links from one of the other Web sites from the set P_S. The local links are ignored. This results in one ranking per other Web site for each Web site. Formally we can specify this ranking as follows. Given the Web link matrix $M \in \mathcal{M}_{\mathbf{P_0}}$ the Web site $s_i \in P_S$ with $D_i = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_i)$ to be ranked and the external Web site $s_j \in P_S$ with $D_j = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_j)$ used as ranking context. We include the case where $i = j$. Then

$$R_{D_{ij}}^{LE} = \Pi_{D_i}(R^{Count}(\Lambda_{D_j}(\Pi_{D_i \cup D_j}(M)))) \in \mathcal{R}_{\mathbf{P_0}}(D_i)$$

3.3 Ranking Aggregation

We illustrate here by using ranking algebra again how the rankings described above can be combined to produce further aggregate rankings. Thus we address several issues discussed in previous sections and demonstrate two points:

1. We show that global document rankings can be determined in a distributed fashion, and thus better scalability can be achieved. Hence ranking documents based on global information not necessarily implies a centralized architecture.
2. We show how local rankings from different sources can be integrated, such that rankings can be made precise and can take advantage of globally unavailable information (e.g. the hidden web) or different ranking contents. Thus a richer set of possible rankings can be made available.

Our goal is to produce a composite ranking for the documents in one of the selected subset of Web sites in P_S from the different rankings that have been described before. The specific way of composition has been chosen with two issues in mind: first, we want to illustrate different possibilities of computing aggregate rankings using the ranking algebra, and second, the resulting composite ranking should exhibit a good ranking quality, which we will evaluate in the experimental section, by comparing to various rankings described in Section 3.2.

The aggregate ranking for a Web site $s_i \in P_S$ with $D_i = \rho_{\mathbf{P_S} \gg \mathbf{P_0}}(s_i)$ is obtained in 3 major steps. First we aggregate the local external rankings by weighting them using the global site ranking. Since for each D_i we can compute a local external ranking $R^{LE}_{D_{ij}}$ relative to D_j, we can obtain a covering vector $RLE^{\mathbf{P_S}}_{\mathbf{P_0}}(D_i)$ over P_S by defining $RLE^{\mathbf{P_S}}_{\mathbf{P_0}}(D_i)(s_j) = R^{LE}_{D_{ij}}$. Using the global site ranking we compose an aggregate local document ranking by using a folding operation

$$R^{LE}_{D_i} = \mathcal{F}^{\mathbf{P_S} \gg \mathbf{P_0}}(RLE^{\mathbf{P_S}}_{\mathbf{P_0}}(D_i), R^{GS}_{P_S})$$

Then we combine this ranking of documents in D_i with the local internal ranking in an ad-hoc fashion, using w_E and w_I as the weights that we give to the external and internal rankings.

$$R^{WA}_{D_i} = \Sigma_2(R^{LE}_{D_i}, R^{LI}_{D_i}, w_E, w_I)$$

In this manner we have now obtained a local ranking for each D_i. We can again use these local rankings to construct a covering vector $RCL^{\mathbf{P_S}}_{\mathbf{P_0}}$ over P_S by

$$RCL^{\mathbf{P_S}}_{\mathbf{P_0}} = R^{WA}_{D_i}$$

Using this covering vector we can obtain a global ranking by applying a folding operation. This time we use the local site ranking to perform the ranking

$$R^{comp}_{D_S} = \mathcal{F}^{\mathbf{P_S} \gg \mathbf{P_0}}(RCL^{\mathbf{P_S}}_{\mathbf{P_0}}, R^{LS}_{P_S})$$

Finally we project the ranking obtained to a Web site

$$R^{comp}_{D_i} = \Pi_{D_i}(R^{comp}_{D_S})$$

This composite ranking we will compare experimentally with some of the basic rankings introduced earlier.

4 Application and Evaluation

4.1 Experimental Setting

In this section we give an illustration of how to apply the ranking algebra in a concrete problem setting. We performed an evaluation of the aggregation approach described above within the EPFL domain which contains about 600 independent Web sites ($\mathbf{P_S}$) identified by their hostnames or IP addresses. We crawled about 270.000 documents found in this domain. Using this document

collection we performed the evaluations using the following approach: we chose two selected Web sites s_1 and s_2, with substantially different characteristics, in particular of substantially different sizes. For those domains we computed the local internal and external rankings. We also put the EPFL portal web server s_h (hostname www.epfl.ch) in the collection, since this is a point where most of the other subdomains are connected to. We consider this subset of documents an excellent knowledge source for information of web site importance. So we have $P_S = \{s_1, s_2, s_h\}$ here. We denote the corresponding document ses D_1, D_2, D_h as in section 3.3.

Then we applied the algebraic aggregation of the rankings obtained in that way, in order to generate a global ranking for the joint domains s_1 and s_2. For local aggregation we chose the values $(w_E, w_I) = (0.8, 0.2)$. This reflects a higher valuation of external links than internal links. One motivation for this choice is the relatively low number of links across subdomains as compared to the number of links within the same subdomain. The resulting aggregate ranking $R^{comp}_{D_1 \cup D_2}$ for the joint domains s_1 and s_2 is then compared to the ranking obtained by extracting from the global ranking $R^{global}_{D_1 \cup D_2}$ computed for the complete EPFL domain (all 270.000 documents) for the joint domains s_1 and s_2. The comparison is performed both qualitatively and quantitatively.

4.2 Qualitative Results

We report on one specific experiment performed in the way described above. The subdomains used are sicwww.epfl.ch, the home of the computing center (280 documents) and sunwww.epfl.ch, the support site for SUN machines (21685 documents). Figure 1 compares the top 25 documents resulting from the two ranking methods. We can observe some substantial differences. In the top 25 list of the aggregate ranking result, the top 4 are obviously more important than the top listed ones from the global PageRank. The 2 obviously important pages "http://sunwww.epfl.ch/" and "http://sicwww.epfl.ch/informatique/" are ranked much lower than some of the software documentation pages. We can assume that this is an effect due to the agglomorate structure of these document collections. These play obviously a much less important role in the composite ranking due to the way of how the ranking is composed from local rankings. It shows that the global page ranking is not necessarily the best possible ranking method. We obtained similar qualitative improvements in the ranking results of other domains.

4.3 Quantitative Comparison

For quantitative comparison of rankings we adopt the Spearman's Footrule. [8]:

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

In the formula, $R_j, j = 0, 1$ are the two ranking vectors to be compared. $R_j(i)$ is the rank of document i.

Doc_ID	Rank_Value	URL	Doc_ID	Rank_Value	URL
4194	0.027078119	http://sunwww.epfl.ch/	1500	0.000121	http://sicwww.epfl.ch/SIC/
82	0.005667276	http://sicwww.epfl.ch/informatique/	10714	6.00E-05	http://sunwww.epfl.ch/Admin/todorov.html
1500	0.002242407	http://sicwww.epfl.ch/SIC/	66021	4.10E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/overview-summary.html
10714	0.001030088	http://sunwww.epfl.ch/Admin/todorov.html	66020	3.40E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/index.html
66021	0.000710622	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/overview-summary.html	168390	3.10E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/allclasses-noframe.html
66020	0.000598713	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/index.html	168389	3.10E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/help-doc.html
168387	0.000539437	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/deprecated-list.html	168388	3.10E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/index-files/index-1.html
168388	0.000539437	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/index-files/index-1.html	168387	3.10E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/deprecated-list.html
168389	0.000539437	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/help-doc.html	66435	3.00E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/overview-summary.html
168390	0.000539437	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/allclasses-noframe.html	65975	2.90E-05	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/overview-summary.html
66435	0.000519678	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/overview-summary.html	405856	2.70E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/relnotes/devdocs-vs-specs.html
65975	0.000500968	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/overview-summary.html	4194	2.40E-05	http://sunwww.epfl.ch/
405856	0.000480859	http://sunwww.epfl.ch/Java/jdk1.4/docs/relnotes/devdocs-vs-specs.html	66434	2.30E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/index.html
66434	0.000404271	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/index.html	3709	2.30E-05	http://sicwww.epfl.ch/SIC/SIC-welcome.html
169526	0.000404097	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/deprecated-list.html	169528	2.30E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/help-doc.html
169527	0.000404097	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/index-files/index-1.html	169527	2.30E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/index-files/index-1.html
169528	0.000404097	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/help-doc.html	169526	2.30E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/deprecated-list.html
65974	0.000389758	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/index.html	65974	2.20E-05	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/index.html
167601	0.000389583	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/deprecated-list.html	167603	2.20E-05	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/help-doc.html
167603	0.000389583	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/help-doc.html	167602	2.20E-05	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/index-files/index-1.html
167602	0.000389409	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/index-files/index-1.html	167601	2.20E-05	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/deprecated-list.html
406487	0.000349716	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/java/lang/Object.html	406487	2.00E-05	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/java/lang/Object.html
409918	0.000267882	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/java/lang/Object.html	26506	2.00E-05	http://sicwww.epfl.ch/SIC/SIC-SII.html
399292	0.000255817	http://sunwww.epfl.ch/Java/jdk1.3/docs/api/java/lang/Object.html	82	1.50E-05	http://sicwww.epfl.ch/informatique/
168214	0.000235534	http://sunwww.epfl.ch/Java/jdk1.4/docs/api/java/lang/String.html	409918	1.50E-05	http://sunwww.epfl.ch/Java/jdk1.2/docs/api/java/lang/Object.html

Fig. 1. URLs ranked 1 to 25 in the composite and global ranking

Since search engines return documents in ranking order, top level documents receive generally much higher attention than documents listed later. To take this into account we customize Spearman's Footrule by a weighting scheme

$$F(R_0, R_1) = \sum_{i=1}^{n} w_0(i)w_1(i)|R_0(i) - R_1(i)| \tag{2}$$

Since 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 for our experiment on the 2 subdomains in Figure 2. The figure shows the ranking distance computed using the adapted Spearman's rule of different rankings with respect to the global ranking $R^{global}_{D_1 \cup D_2}$ for varying values of T. Besides of 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, i.e. $R^{intermediate}_{D_1 \cup D_2}$. This ranking thus uses exactly the same information that is available to the computation of the aggregate ranking, i.e., the documents in the set P_S. The "tinyset" ranking is the ranking obtained by selecting exactly all documents that are ranked by the aggregate ranking and applying PageRank to them, which is exactly $R^{LI}_{D_1 \cup D_2}$. In addition, we included for calibration a randomly generated ranking. The results are shown in Figure 2.

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, 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

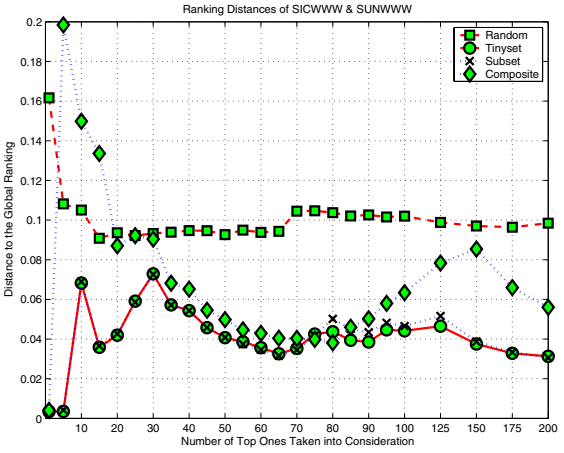

Fig. 2. Ranking Distances of SICWWW & SUNWWW

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.

Due to limit of space, we only show the main results here. More results can be found in a longer version of this paper at http://lsirwww.epfl.ch/.

4.4 Summary

From the comparison and analysis, we find that with our ranking algebra, the ranking result has been improved in two important aspects: firstly, default important pages (for example the department home) are levered to the rank that they deserve; secondly, the reinforcing effect of some agglomerate pages is defeated to a satisfactory degree.

In short, our results making use only of local information approximate the result of PageRank based on global information very well and in some cases appear to be even better with respect to importance of documents. We see this work as a first step towards a completely decentralized P2P-based search engine that offers meaningful and efficient rankings.

5 Future Work

By introducing a ranking algebra we made a first step towards an operational framework for manipulating and composing rankings. An obvious development is to determine and exploit algebraic equivalences in order to find alternative plans for computing rankings in a distributed environment most efficiently.

Having the possibility to consider different contexts for rankings, an interesting approach is to use information obtained from user interactions in order to obtain information on the relevance of documents. This kind of local feedback could greatly enhance the quality of local rankings that could be used in the framework that allows to integrate different local rankings.

Another specific context of which composite rankings can take advantage of are so-called hub sites. These are special Web sites that provide a directory function by pointing to many relevant (authority) sites. They would be particularly useful to provide Web site rankings used to fold multiple local rankings, as illustrated in our examples.

Acknowledgements. We highly appreciate the discussions with Prof. Ling Liu, Prof. Calton Pu, and their student Todd Miller from Georgia Institute of Technology during their academic sojourn at LSIR (Distributed Information Systems Laboratory), Swiss Federal Institute of Technology, Lausanne.

References

1. Sergey Brin, Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 2000.
2. Larry Page, Sergey Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", 1998.
3. Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", 1998.
4. Danny Sullivan, "New AllTheWeb.com Goes Live", http://searchenginewatch.com/sereport/01/08-alltheweb.html, 2001.
5. Chris Sherman, "It's Fresher at FAST", http://www.searchenginewatch.com/searchday/01/sd0725-fast.html, 2001.
6. Mitch Wagner, "Google Bets The Farm On Linux", http://www.internetwk.com/lead/lead060100.htm, 2000.
7. UC Berkeley SIMS, "How Much Information – Internet Summary", http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html, 2000.
8. Keith A. Baggerly, "Visual Estimation of Structure in Ranked Data", PhD thesis, Rice University, 1995.
9. K. Bharat, M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment", in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 104–111, Melbourne, Australia, August 1998. ACM Press, New York.
10. Danny Sullivan, "Google Adds More "Fresh" Pages, Changes Robots.txt & 403 Errors, Gains iWon", http://searchenginewatch.com/sereport/02/08-google.html, Aug. 5, 2002.