Authoritative Sources in a Hyperlinked Environment *

Jon M. Kleinberg†

Abstract

The link structure of a hypermedia environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. Versions of this principle have been studied in the hypertext research community and (in a context predating hypermedia) through journal citation analysis in the field of bibliometrics. In this work we develop a set of algorithmic techniques for studying the link structure of hypermedia environments, and we report on experiments that demonstrate their effectiveness in several contexts on the World Wide Web. One issue that our methods address is that of identifying WWW pages that are "authoritative sources" on broad search topics; this problem arises, for example, when trying to select a high-quality set of relevant "representative pages" from among a large collection of pages retrieved by a text-based search method. We propose and test an algorithmic formulation of the notion of authority, based on a method for locating dense bipartite *communities* in the link structure. Our formulation has an interesting interpretation in terms of the eigenvectors of certain matrices associated with the link graph; this motivates additional heuristics for clustering and for computing a type of link-based similarity among hyperlinked documents.

1 Introduction

The link structure of a hypermedia environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. Versions of this principle have been studied in the hypertext research community [2, 8, 16, 25] and (in a context predating hypermedia) through journal citation analysis in the field of bibliometrics [26]. But for the problem of searching in hyperlinked environments such as the World Wide Web, it is clear from the prevalent techniques that the information inherent in the links has yet to be fully exploited. In this work we develop a new method for automatically extracting certain types of information about a hypermedia environment from its link structure, and we report on experiments that demonstrate its effectiveness in a variety of contexts on the WWW.

Our methods seem to apply fairly broadly, to structures that are implicitly, as well as explicitly, linked. In the present context, we focus on the development and testing of algorithms for searching in hypermedia, particularly on the World Wide Web. We will show some interesting connections between our algorithms and the spectral properties of certain matrices derived from the link structure of the underlying environment; it is through these connections that we will be able to develop some insight into their behavior, and to prove certain convergence properties.

Searching, in the setting of the WWW and the present work, could be defined as the process of discovering pages that are relevant to a given query. The *quality* of a search method necessarily requires human evaluation, to make concrete the various loaded terms in the previous sentence. We begin from the observation that improving the quality of search methods on the WWW is, at the present time, a rich and interesting problem that is in many ways orthogonal to concerns of algorithmic efficiency and storage. In particular, consider that current search engines typically index a large fraction of the WWW and respond on the order of seconds. Although there would be considerable utility in a search tool whose response time was on the order of minutes, provided that the results were of significantly greater value to a user, it has typically been very hard to say *what* such a search tool should be computing with this extra time. Clearly we are lacking an objective function that is both concretely defined *and* corresponds to human notions of quality.

Our work is centered around this issue of improving the quality of search results; we are seeking new methodologies for searching in large hyperlinked environments, rather than focusing on efficient implementations of existing techniques. In particular, the initial emphasis of our work is to *define*, by algorithmic means, a novel type of quality measure that we refer to as the *authority* of a document in hypermedia; a highly authoritative docu-

*A full version of this paper is available from the author, at http://www.cs.cornell.edu/home/kleinber/, and as IBM Research Report RJ 10076(91892) May 1997.

†Department of Computer Science, Cornell University, Ithaca NY 14853. Email: kleinber@cs.cornell.edu. This work was performed in large part while on leave at the IBM Almaden Research Center, San Jose CA 95120. The author is currently supported by an Alfred P. Sloan Research Fellowship and by NSF Faculty Early Career Development Award CCR-9701399.

ment intuitively represents a high-quality response to a broad user query. Our algorithmic definition naturally yields an efficient means to compute this authority measure; moreover, an analysis of our method in terms of eigenvectors motivates additional useful heuristics that would have been difficult to formulate without appealing to spectral methods. We feel that the interplay between this spectral analysis and the motivation for our heuristics is one of the interesting features of this work.

Queries and Authoritative Sources. We view *searching* as beginning from a user-supplied *query*. It seems best not to take too unified a view of the notion of a *query*: there are many possible types of queries that one might wish to pose, and it is likely that the proper handling of each type requires a different set of techniques. Consider, for example, the following types of queries.

- *Broad-topic queries.* E.g., "Find information about web browsers."
- *Specific queries.* E.g., "Has the WWW Consortium endorsed the HTML 3.2 specification?"
- *Similar-page queries.* E.g., "Find pages 'similar' to `www.lcs.mit.edu`."

Concentrating on just the first two types of queries for now, we see that they present very different sorts of obstacles. The difficulty in handling *specific queries* is centered, roughly, around what could be called the *Scarcity Problem*: there are very few pages that contain the required information, and it is difficult to determine the identity of these pages. Much classical work in information retrieval has focused on this type of problem.

For *broad-topic queries*, on the other hand, one could easily expect to find many thousand relevant pages in an environment such as the WWW; such a set of pages might be generated by variants of term-matching (e.g. one enters a string such as "web browsers," "Gates," or "censorship" into a search engine such as AltaVista [4]), or by more sophisticated means. Thus, there is not an issue of scarcity here. Instead, the fundamental difficulty lies in what could be called the *Abundance Problem*: *The number of pages that could reasonably be returned as "relevant" is far too large for a human user to digest.* Thus, to provide effective methods for automated search under these constraints, one does not necessarily need stronger versions of classical information retrieval notions such as relevance; rather one needs a method of providing a user, from a large set of relevant pages, a small collection of the most "authoritative" or "definitive" ones.

Our work here originates from these issues raised by the Abundance Problem, and the problem of discovering the most authoritative pages in a large hyperlinked environment. The problem is particularly interesting in that much of its complexity has nothing to do with the "search" component; rather, we face the dilemma that in order to search for authoritative documents, one must first formulate a concrete means of *recognizing* them. Unfortunately, "authority" is perhaps an even more nebulous concept than "relevance," again highly subject to human judgment; and our algorithmic framework must take this into account.

It is here that we bring the notion of links into the picture. We claim that an environment such as the WWW is explicitly annotated with precisely the type of human judgment that we need in order to formulate a notion of authority. Specifically, the creation of a link in the WWW represents a concrete indication of the following type of judgment: the creator of page p, by including a link to page q, has in some measure *conferred authority* on q. Of course, this notion is clouded by the fact that links are created for a wide variety of reasons, many of which have nothing to do with the conferral of authority. Thus we are faced with the following problem: given the vast size of the underlying environment, can we synthesize the unreliable information contained in the presence of individual links in a way that provides a set of *authoritative* pages *relevant* to an initial query?

A Crude Approximation. Naturally, we first examine the simplest implementation of the above idea: if the presence of links is an indication of authority, can one simply use the *in-degree* of a page as a measure of its *authority*?

There are several variants of this idea, and none works very well. First of all, since we are looking for pages that are relevant as well as authoritative, we must specify the subgraph of the WWW in which we are computing the in-degree. As an example, consider the query `"java"`, a string contained in more than two million pages on the WWW.

(1) One approach is to define a *root set* S as follows. For a number k (say 200), we define S to be the top k pages indexed by AltaVista (or some other term-based search engine); we then rank pages according to their in-degree in the subgraph induced by S. There are two severe problems with this approach. First, this subgraph typically has very few edges; a large fraction (if not most) of the nodes will be isolated. Second, the root set S, for any reasonable value of k, omits most of the pages that one would normally consider authoritative for the query `"java"`; they are not ranked highly enough by AltaVista's scoring function.

(2) A more reasonable approach is the following. We start from the same root set S, and we then grow it to a larger *base set* T, consisting of all pages that either belong to S, point to a page in S, or are pointed to by a

http://www.gamelan.com	*Gamelan*
http://java.sun.com	*JavaSoft Home Page*
http://getawaynet.com/index.html	*GetAwayNet Home Page*
http://getawaynet.com/VacationNetwork/vnetwork.html	*Caribbean Vacation Network Home Page - 1-800-423-4095*
http://www.sn.no/~espeset	*Java Programming*
http://www.net4u.ch/net4u/ger/index-ger.html	*Net4U Tielseite*
http://www.net4u.ch/net4u/eng/index-eng.html	*Net4U Home Page*
http://www.amazon.com/exec/obidos/stores/jollyrog	*Welcome to Amazon.com Books! Earth's Biggest Bookstore*

Figure 1: Pages for "java", sorted purely by in-degree.

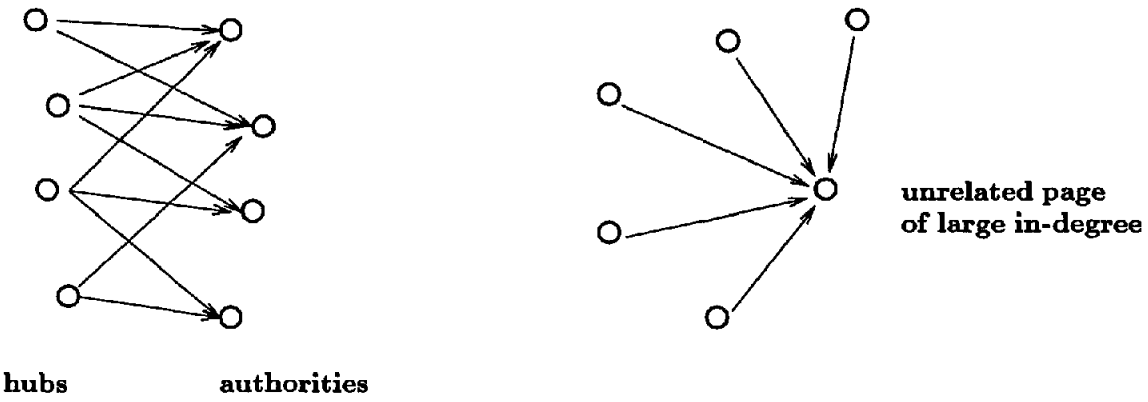

Figure 2: A dense community of hubs and authorities.

page in S. (To prevent the size of T from exploding, we arbitrarily truncate the in-degree of pages in S to some upper bound d.) We then rank pages by their in-degree in the subgraph induced by T. The base set T has two attractive features: it is extremely likely to contain many authoritative pages on the topic (since they will be pointed to by pages in S); and for sufficiently broad queries, it will still generally be "rich" in relevant pages.

Unfortunately, the results are still far from satisfying; we list the top eight pages for "java" in Figure 1. The list in this figure has several features that are worth comment, as they arise generally when ranking the results of a broad-topic query purely by in-degree. A promising feature of the list is that the first two pages should certainly be viewed as "good" answers. However, five of the six other pages are not relevant to the original query — they are advertisements for Carribean vacations, pages for a Swiss Internet consulting company, and the home page of Amazon Books; and the sixth page is an advertisement for a single book on Java programming. Although these pages all have large in-degree, they lack any thematic unity; we have not achieved the goal of providing pages that are authoritative *and* relevant.

Our Approach. It is tempting to conclude that the only way to preserve relevance while looking for authoritative pages is to make use of the text of pages, as current search engines do. However, this is a tricky issue. For example, all but the last of the eight pages listed above contain the string "java", most of them multiple times. Moreover, to use our three earlier examples, it would natural to want to find Netscape's home page for the query "web browsers", Microsoft's home page for the query "Gates", and the Electronic Frontier Foundation's home page for the query "censorship". Unfortunately, none of these pages contain the respective query term.

While much work has gone into text-based methods for circumventing these relevance-related difficulties (e.g. latent semantic indexing [6] and a range of other clustering techniques in information retrieval), our goal here is to understand how much can be accomplished by focusing on link structures, for finding pages that are simultaneously authoritative and relevant. We will see that quite striking results can be achieved while making essentially no use of text whatsoever.

Our approach proceeds essentially as follows. From an initial query, we form the *base set* T defined previously; this set has the useful features discussed above. We now make the following observation. Authoritative pages relevant to the initial query should not only have large in-degree; since they are all authorities on a common topic, there should also be considerable overlap in the *sets* of pages that point to them. Thus, in addition to highly authoritative pages, we expect to find what could be called *hub pages*: these are pages that

have links to multiple relevant authoritative pages. It is these hub pages that "pull together" authorities on a common topic, and allow us to throw out unrelated pages of large in-degree.

Hubs and authorities exhibit what could be called a *mutually reinforcing relationship*: a good *hub* is a page that points to many good authorities; a good *authority* is a page that is pointed to by many good hubs. Clearly, if we wish to identify hubs and authorities within the base set T, we need a method for breaking this circularity. In Section 2, we describe our basic algorithm for this task — a method that iteratively propagates "authority weight" and "hub weight" across links of the web graph, converging simultaneously to steady states for both types of weights. The final output of our algorithm, derived from an equilibrium set of weights, is a pair of sets (X, Y), where X is a small set of authorities and Y is a small set of hubs; this is the desired small set of "high-quality" pages that can be returned in response to a user query. We refer to the pair of sets (X, Y) as a *community* of hubs and authorities, characterized by their mutually reinforcing relationship; one can picture this pair as the two sides of a dense directed bipartite subgraph of the base set T, with the hubs linking densely to the authorities. (A skeletal example is depicted in Figure 2; in reality, of course, the picture is not nearly this clean.)

Thus our central claim is that authoritative pages can be identified as belonging to dense bipartite *communities* in the link graph of the WWW, via the algorithm described in the following section. A valuable feature of our techniques here is that they are robust in several respects. Although we may be dealing with query topics with up to several million relevant pages, we are arriving at quite reliable estimates of authoritative pages by examining only the few thousand pages in the base set T. This clearly has to do with the notion of "authority" — we are guided by the intuition that one can find authoritative pages starting from almost *any* small root set S, provided that the WWW contains a sufficient number of relevant pages on the query topic. Thus, the effectiveness of our technique appears not to be hampered by the phenomenal rate of growth of the WWW; indeed, there are indications that it produces more reliable results for search topics with greater numbers of relevant pages.

Overview. In Section 2, we describe the basic method and show examples of its behavior for finding authoritative pages. We show that our iterative weight-assignment algorithm can be analyzed as an eigenvector computation on a pair of matrices derived from the Web graph; among other things, this allows us to prove its convergence.

For many query topics, the base set T may contain *multiple* dense communities of hubs and authorities, which link sparsely, if at all, to one another. This may arise for several reasons: there could "on-topic" versus "off-topic" communities (e.g. a community of pages on java together with a smaller community on Caribbean vacations); the query term could have multiple meanings or uses in different settings (e.g. "`jaguar`"); or the query term could refer to a polarized issue involving groups that will not link to one another (e.g. "`abortion`"). The spectral interpretation of our algorithm provides us with a natural way to discover many of these additional communities: they correspond to coordinates of large absolute value in non-principal eigenvectors. (For this reason, we will at times refer to them as *non-principal communities*.) We discuss this issue in Section 3. The heuristic intuition behind this approach is analogous to the spectral partitioning of undirected graphs (e.g. [5, 7, 22]); however, it is important to note that what we are doing here is not simply a spectral partitioning of the Web graph. In particular, we are studying non-principal eigenvectors of symmetric matrices derived from the (asymmetric) adjacency matrix of the base set; and the structures we find are dense directed bipartite subgraphs, rather than simply sparse partitions of the node set.

In Section 4, we apply our method to the problem of *similar-page queries*: given a page p of large in-degree, we construct a base set in the neighborhood of p and determine the good authorities. This results in an interesting notion of page similarity, defined by the link structure.

Our experiments with the technique have been directed primarily at trying to understand the nature and quality of the output that is produced, and the extent to which it corresponds to human judgment. The difficulty in assessing the results of our algorithm is clear. We are attempting to define a new measure, in a domain that is itself quite new. In evaluating output that requires the judgment of a user, one typically makes use of human-annotated benchmarks; unfortunately, such benchmarks are not directly available in the present setting. Given this, we adopt the following multi-pronged approach to evaluating the output. First, we claim that an element of *res ipsa loquitur* applies: we feel that many of our results are quite striking at a fairly obvious level, and for a variety of reasons would be hard to produce using standard search methods currently available on the WWW. Second, there is a sense in which we *can* compare the technique to existing human-constructed benchmarks: there are a number of *searchable hierarchies* on the WWW, such as YAHOO[27], Galaxy [23], and the distributed WWW Virtual Library [24]. These hi-

erarchies provide lists of authoritative pages, compiled by human moderators, for many standard search topics. In this way, they represent high-quality hub pages, and can be used for comparison with our automated method. Third, we can observe that for broad search topics, the communities discovered by our method have a *robust* identity — that is, using several different root sets on the same "topic" produces similar communities of hubs and authorities. For reasons of space, we will not be able to go into detail about these latter points in this version; we refer the reader to the full paper for a discussion of these issues.

In the full version, we also discuss experiments aimed at assessing the principal limitation of our technique — that a query topic must be sufficiently "broad" in order for our method to produce reliable sets of hubs and authorities. In particular, we study the following recurring phenomenon: when the query defines a topic that is relatively specific, the principal community of hubs and authorities is often relevant to a generalization of the query topic, rather than to the initial topic itself. It is fairly clear why this should happen: our algorithm is designed to locate the "densest" community of hubs and authorities in the base set T, without regard to the initial query, and hence it will favor topics that have large representation in the vicinity of a root set derived from the query. We refer to this process as *diffusion*; the focus of the query has "diffused" to a generalization of the original topic. We show in the full paper that non-principal eigenvectors can be a very effective way to produce relevant results even in the face of this phenomenon; in particular, a community relevant to the (specific) initial query often exists and corresponds to one of the non-principal eigenvectors.

We noted at the outset that our primary interest was in studying the *quality* of search methods for hypermedia, even at the cost of algorithmic efficiency. However, the method we discuss here ultimately turns out to be relatively efficient. In particular, the main computational step of the algorithm can be reduced to a singular value decomposition of a sparse matrix with several thousand non-zeroes — a task for which highly optimized code is available. Moreover, as we have already discussed, our method produces meaningful results using only a small fragment of the WWW for each query. We will not be focusing further on the question of efficiency in this paper. (Note that since we did not have the WWW indexed locally while performing these experiments, the time required simply to fetch the `html` source of several hundred or several thousand pages, so as to construct the base set, proved to be a greater bottleneck. This is an issue that is largely separate from the computational requirements of our algorithm. For example, in preliminary experiments on a *local* corpus of two million U.S. patents [12], these problems associated with processing the documents did not arise.)

Related Work on Link Structures. Methodologically, our work has connections to the area of ***bibliometrics*** [26] — the study of written documents and their citation structure. Bibliometrics has developed citation-based measures of document similarity, such as ***bibliographic coupling*** [13] and ***co-citation*** [20]; these have been used in the context of document clustering (see e.g. [18, 19]). There has also been work in bibliometrics on using citation counts to assess the "impact" of scientific journals; this problem is more closely related to the issue we are considering here, though the underlying techniques are different. The classic work in this area is that of Garfield [9]; see also e.g. [10, 17].

There has been work on using hyperlinks for clustering and searching in the hypertext research community. Rivlin, Botafogo, and Schneiderman [2, 16], and Weiss et al. [25], use local graph-theoretic notions to determine similarity and clustering information about documents in hypertext. Frisse [8] describes a link-based method for enhancing relevant judgments in tree-structured hypertext.

More recently, Pirolli, Pitkow, and Rao [15] have used a combination of link topology and textual similarity to group together and categorize pages on the WWW. Arocena, Mendelzon, and Mihaila [1] and Spertus [21] have described frameworks for constructing WWW queries from a combination of term-matching and link-based predicates. Page [14] has developed a method for assigning a universal "rank" to each page on the WWW, so that subsequent user searches can be focused on highly ranked pages; the rank of a page is based on a weight-propagation algorithm that corresponds roughly to simulating a short random walk on the directed link graph of the WWW. Finally, Carrière and Kazman [3] propose a link-based method for visualizing and ranking the results of queries returned by WWW search engines. Their method is based on sorting pages by their degree in the WWW graph; but they do not make use of the directionality of links, and do not work directly at producing multiple dense clusters of pages in the link structure.

2 The Method

We first give a description of the basic algorithm. The algorithm can be run on an arbitrary set of hyperlinked pages, and we represent such a set as a directed graph $G = (V, E)$ in the natural way: V consists of the n pages in the environment, and a directed edge $(p, q) \in E$ indicates the presence of a link from p to q. In the applications discussed below, G will typically be the

subgraph induced on our *base set* T.

We associate a non-negative *authority weight* x_p and a non-negative *hub weight* y_p with each page $p \in V$. We maintain the invariant that the weights of each type are normalized so their squares sum to 1: $\sum_{p \in V} x_p^2 = 1$, and $\sum_{p \in V} y_p^2 = 1$. We view pages with larger x- and y-values as being "better" authorities and hubs respectively.

The Iterative Algorithm. Recall the mutually reinforcing relationship between hubs and authorities. Numerically, it is natural to express this as follows: if p points to many pages with large x-values, then it should receive a large y-value; and if p is pointed to by many pages with large y-values, then it should receive a large x-value. This motivates the definition of two operations on the weights, which we denote by $\mathcal{I}$ and $\mathcal{O}$. Given weights $\{x_p\}$, $\{y_p\}$, the $\mathcal{I}$ operation updates the x-weights as follows.

$$x_p \leftarrow \sum_{q:(q,p) \in E} y_q.$$

The $\mathcal{O}$ operation updates the y-weights as follows.

$$y_p \leftarrow \sum_{q:(p,q) \in E} x_q.$$

Thus $\mathcal{I}$ and $\mathcal{O}$ are the basic means by which hubs and authorities reinforce one another.

Now, to find the desired "equilibrium" values for x and y, one can apply the $\mathcal{I}$ and $\mathcal{O}$ operations in an alternating fashion, and see whether a fixed point is reached. Indeed, we can now state a version of our basic algorithm.

```
Let z denote the vector (1,1,1,...,1).
Initially set x ← z; y ← z.
For i = 1,2,3,...
    Apply the I operation
    Apply the O operation
    Normalize x and y
The sequence of (x,y) pairs produced
  converges to a limit (x*,y*) (Thm 2.1).
Return (x*,y*) as authority/hub weights.
```

The basic convergence result is not difficult to prove; we develop it here. For a pair of authority/hub weight vectors (x, y), let $(\mathcal{OI})(x, y)$ denote the result of applying the $\mathcal{I}$ operation followed by the $\mathcal{O}$ operation (and normalizing the vectors obtained). Let $(\mathcal{OI})^n(x, y)$ denote the result of doing this n times. Finally, as above, let z denote the vector in $\mathbf{R}^n$ in which each coordinate is equal to 1, and let $(x_n, y_n) = (\mathcal{OI})^n(z, z)$.

For the proof, we need the following additional notions. For an $n \times n$ symmetric matrix M, let $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$ denote the eigenvalues of M (all of which are real), indexed in order of decreasing absolute value. Let $\omega_i(M)$ denote the eigenvector associated with λ_i. For the sake of simplicity, we will make the following technical assumption about all the matrices we deal with:

(†) $|\lambda_1(M)| > |\lambda_2(M)|$.

When this assumption holds, we refer to $\omega_1(M)$ as the *principal eigenvector*, and all other $\omega_i(M)$ as *non-principal eigenvectors*. When the assumption does not hold, the analysis becomes less clean, but it is not affected in any substantial way.

THEOREM 2.1. *The sequences $x_1, x_2, x_3, \ldots$ and $y_1, y_2, y_3, \ldots$ converge (to limits x^* and y^* respectively).*

Proof. Write $V = \{p_1, p_2, \ldots, p_n\}$, and let A denote the *adjacency matrix* of the graph G; the $(i, j)^{\text{th}}$ entry of A is equal to 1 if (p_i, p_j) is an edge of G, and is equal to 0 otherwise. One easily verifies that the $\mathcal{I}$ and $\mathcal{O}$ operations can be written $x \leftarrow A^T y$ and $y \leftarrow Ax$ respectively. Thus x_n is the unit vector in the direction of $(A^T A)^{n-1} A^T z$, and y_n is the unit vector in the direction of $(AA^T)^n z$.

Now, a standard result of linear algebra (see e.g. [11]) states that if M is a symmetric $n \times n$ matrix, and v is a vector not orthogonal to the principal eigenvector $\omega_1(M)$, then the unit vector in the direction of $M^n v$ converges to $\omega_1(M)$ as n increases without bound. Also (as a corollary), if M has only non-negative entries, then the principal eigenvector of M has only non-negative entries.

Consequently, z is not orthogonal to $\omega_1(AA^T)$, and hence the sequence $\{y_n\}$ converges to a limit y^*. Similarly, one can show that if $\lambda_1(A^T A) \neq 0$ (as dictated by Assumption (†)), then $A^T z$ is not orthogonal to $\omega_1(A^T A)$. It follows that the sequence $\{x_n\}$ converges to a limit x^*. ∎

The proof of Theorem 2.1 yields the following additional result (in the above notation).

THEOREM 2.2. *(Subject to Assumption (†).) x^* is the principal eigenvector of $A^T A$, and y^* is the principal eigenvector of AA^T.*

As indicated above, the output of our process from the user's point of view would a pair of sets (X, Y): the c pages with the largest x^*-values and the c pages with the largest y^*-values, for a small constant c. This represents the algorithm's estimate of the strongest authorities and hubs.

Theorem 2.2 directly allows one to develop methods for computing x^* and y^* that are more efficient than the iteration described above. We have stuck to the above exposition for two reasons. First, it emphasizes

("web browsers") Authorities	
.225 http://www.ncsa.uiuc.edu/SDG/Software/WinMosaic/HomePage.html	*NCSA Windows Mosaic Home Page*
.202 http://home.mcom.com/home/welcome.html	*Welcome to Netscape*
.196 http://galaxy.einet.net/EINet/EINet.html	*TradeWave Corporation*
.188 http://www.interport.net/slipknot/slipknot.html	*.... SlipKnot Home Page ....*
.188 http://galaxy.einet.net/EINet/WinWeb/WinWebHome.html	*winWeb and MacWeb*
.185 http://www.microsoft.com/ie/	*Microsoft Internet Explorer*

Figure 3: The query (`"web browsers"`).

(java) Authorities	
.328 http://www.gamelan.com/	*Gamelan*
.251 http://java.sun.com/	*JavaSoft Home Page*
.190 http://www.digitalfocus.com/digitalfocus/faq/howdoi.html	*The Java Developer: How Do I...*
.190 http://lightyear.ncsa.uiuc.edu/~srp/java/javabooks.html	*The Java Book Pages*
.183 http://sunsite.unc.edu/javafaq/javafaq.html	*comp.lang.java FAQ*
(+censorship +net) Authorities	
.421 http://www.eff.org/	*EFFweb - The Electronic Frontier Foundation*
.394 http://www.eff.org/blueribbon.html	*The Blue Ribbon Campaign for Online Free Speech*
.390 http://www.cdt.org/	*The Center for Democracy and Technology*
.374 http://www.vtw.org/	*Voters Telecommunications Watch*
.291 http://www.aclu.org/	*ACLU: American Civil Liberties Union*
(Gates) Authorities	
.643 http://www.roadahead.com/	*Bill Gates: The Road Ahead*
.458 http://www.microsoft.com/	*Welcome to Microsoft*
.440 http://www.microsoft.com/corpinfo/bill-g.htm	

Figure 4: Results for three sample queries.

the underlying motivation for our approach in terms of the reinforcing $\mathcal{I}$ and $\mathcal{O}$ operations. Second, one does not have to run the above process of iterated $\mathcal{I}/\mathcal{O}$ operations to convergence; one can instead compute weights $\{x_p\}$ and $\{y_p\}$ by starting from the "flat" vector z and performing a fixed bounded number of $\mathcal{I}$ and $\mathcal{O}$ operations. In many of our experiments, even using a small number of iterations gives good results.

Computing Hubs and Authorities. Given a user-supplied query string, our overall method for finding authoritative pages is now the following. (We require parameters k and d, which in our experiments we assign default values of 200 and 50 respectively.)

(1) We supply the query string to a term-based search engine such as AltaVista; this returns a set S of k pages, which we refer to as the *root set*.

(2) We then enlarge the root set to a *base set* T. Recall from the introduction that T consists of all pages that belong to S, point to a page in S, or are pointed to by a page in S — with the restriction that we allow a single page in S to bring at most d pages pointing to it into T. This latter point is crucial since a number of WWW pages have an in-degree in the hundreds of thousands, and we want to keep T reasonably small.

(3) We define the graph G to be the subgraph of the WWW induced on the set T. We delete from G all links between pages with the same domain name; we find that such *intrinsic links* very frequently serve as navigational aids and have little to do with conferring authority. We then run the iterative algorithm on the graph G, obtaining sets of hubs and authorities.

Basic Results. We now give some sample results, using the queries discussed in the introduction. For (`"web browsers"`) we obtain the results in Figure 3; the top six authorities found by the algorithm are listed. Each line in Figure 3 gives the x-value of a page, followed by its URL, and then its `html` title. Thus, we note that the above set of authorities includes home pages for NCSA Mosaic, Netscape, and Microsoft Internet Explorer (as well as home pages for other browser manufacturers). It is also worth noting that no pages from any of these three domains were included in the initial root set S provided by AltaVista.

For the queries (`java`), (`+censorship +net`), and (`Gates`), the top authorities found by the algorithm are listed in Figure 4. (For the second query, the syntax indicates that both words must appear in the initial pages returned by the search engine.) Among all these pages, the only one which occurred in the corresponding root set S was `www.roadahead.com/`, under the query (`Gates`); it was ranked 123rd by AltaVista. This is natural in view of the fact that many of these pages do

(abortion) Authorities: 2nd non-principal vector, positive end	
.321 http://www.caral.org/abortion.html	*Abortion and Reproductive Rights Internet Resources*
.219 http://www.plannedparenthood.org/	*Welcome to Planned Parenthood*
.195 http://www.gynpages.com/	*Abortion Clinics OnLine*
.172 http://www.oneworld.org/ippf/	*IPPF Home Page*
.162 http://www.prochoice.org/naf/	*The National Abortion Federation*
.161 http://www.lm.com/~lmann/feminist/abortion.html	

(abortion) Authorities: 2nd non-principal vector, negative end	
-.197 http://www.awinc.com/partners/bc/commpass/lifenet/lifenet.htm	*LifeWEB*
-.169 http://www.worldvillage.com/wv/square/chapel/xwalk/html/peter.htm	*Healing after Abortion*
-.164 http://www.nebula.net/~maeve/lifelink.html	
-.150 http://members.aol.com/pladvocate/	*The Pro-Life Advocate*
-.144 http://www.clark.net/pub/jeffd/factbot.html	*The Right Side of the Web*
-.144 http://www.catholic.net/HyperNews/get/abortion.html	

Figure 5: Positive and negative entries of the second non-principal eigenvector for (abortion).

not contain any occurrences of the initial query term.

All of the above topics were sufficiently "broad" that the principal set of authorities (and hubs) were relevant to the query. As mentioned in the introduction, we address the notion of *diffusion* — when the search topic is not sufficiently broad — in the full version of the paper.

3 Multiple Communities

The basic search method described above is, in a sense, finding the *densest* community of hubs and authorities in the base set T. However, there are a number of settings in which one might be interested in finding several distinct communities among the set of pages. There may be several dense communities, only one of which is relevant to the query topic. Alternately, there may be several communities, all of them relevant, but well-separated from one another in the graph on T for a variety of possible reasons. For example,

(1) The string may have several different meanings. E.g. (jaguar).

(2) The string may arise as a term in the context of multiple technical communities. E.g. ("randomized algorithms").

(3) The string may refer to a highly polarized issue, involving groups that are not likely to link to one another. E.g. (abortion).

The non-principal eigenvectors of the matrices A^TA and AA^T provide us with a natural way to extract multiple communities of hubs and authorities from the base set T. We first note the following simple fact.

THEOREM 3.1. *AA^T and A^TA have the same multiset of eigenvalues, and their eigenvectors can be chosen so that $\omega_i(AA^T) = A\omega_i(A^TA)$.*

Thus, each pair of eigenvectors $x_i^* = \omega_i(A^TA)$, $y_i^* = \omega_i(AA^T)$, related as in Theorem 3.1, has the following property: applying an $\mathcal{I}$ operation to (x_i^*, y_i^*) keeps the x-weights parallel to x_i^*, and applying an $\mathcal{O}$ operation to (x_i^*, y_i^*) keeps the y-weights parallel to y_i^*. Hence, each pair of weights (x_i^*, y_i^*) has precisely the *mutually reinforcing relationship* that we are seeking in authority/hub pairs. Moreover, applying $(\mathcal{IO})$ (resp. $(\mathcal{OI})$) multiplies the magnitude of x_i^* (resp. y_i^*) by a factor of $|\lambda_i|$; thus $|\lambda_i|$ gives precisely the extent to which the hub weights y_i^* and authority weights x_i^* *reinforce* one another.

Now, the non-principal eigenvectors have both positive and negative entries. Hence each pair (x_i^*, y_i^*) provides us with two communities of authorities and hubs: (X_i^+, Y_i^+), consisting of the c pages with the most positive coordinates in x_i^* and y_i^*; and (X_i^-, Y_i^-), consisting of the c pages with the most negative coordinates in x_i^* and y_i^*. Such communities have the same intuitive meaning as those produced in the previous section, although the algorithm to find them — based on non-principal eigenvectors — is certainly less intuitive than the method of iterated $\mathcal{I}$ and $\mathcal{O}$ operations. (It is possible to modify that method by adding a Gram-Schmidt step so as to obtain these additional communities.) Note that communities associated with eigenvectors of larger absolute value will tend to have more intuitive meaning, since they are "denser" as subgraphs in the link structure; we will sometimes refer to this notion as the *strength* of a community.

Another interesting feature of the communities derived from non-principal eigenvectors is the following. Drawing on the heuristic intuition underlying *spectral graph partitioning* [5, 7, 22], one expects pairs of communities (X_i^+, Y_i^+) and (X_i^-, Y_i^-) associated with the *same* eigenvector to be very sparsely connected in the underlying graph. In some cases, this sparse linkage can have meaning in the context of the query topic.

One case in which the meaning of this separation is particularly striking is for the query (abortion). The natural question is whether one of the non-

(www.honda.com) Authorities: principal eigenvector	
.202 http://www.toyota.com/	*Welcome to @Toyota*
.199 http://www.honda.com/	*Honda*
.192 http://www.ford.com/	*Ford Motor Company*
.173 http://www.bmwusa.com/	*BMW of North America, Inc.*
.162 http://www.volvocars.com/	*VOLVO*
.158 http://www.saturncars.com/	*Welcome to the Saturn Web Site*
.155 http://www.nissanmotors.com/	*NISSAN - ENJOY THE RIDE*
.145 http://www.audi.com/	*Audi Homepage*
.139 http://www.4adodge.com/	*1997 Dodge Site*
.136 http://www.chryslercars.com/	*Welcome to Chrysler*

Figure 6: Top pages "similar" to `www.honda.com`.

(www.nytimes.com) Authorities: 1st non-principal vector, positive end		
.111	http://www.microsoft.com/	*Welcome to Microsoft*
.110	http://www.ibm.com/	*IBM Corporation*
.101	http://www.apple.com/	*Apple Computer*
.100	http://www.hp.com/	*Welcome to Hewlett-Packard*
.098	http://www.sun.com/	*Sun Microsystems*
.097	http://www.intel.com/	*Welcome to Intel*
.097	http://www.novell.com/	*Novell World Wide: Corporate Home Page*
.087	http://www.ustreas.gov/	*Welcome To The Department of Treasury*
.084	http://www.compuserve.com/	*Welcome to CompuServe*
.081	http://www.lcs.mit.edu/	*MIT Lab for Computer Science Web Page*

(www.nytimes.com) Authorities: 1st non-principal vector, negative end		
-.220	http://www.nytimes.com/	*The New York Times on the Web*
-.169	http://www.usatoday.com/	*USA TODAY*
-.138	http://www.cnn.com/	*CNN Interactive*
-.091	http://www.sjmercury.com/	*Mercury Center*
-.080	http://www.chicago.tribune.com/	*The Chicago Tribune*
-.076	http://www.washingtonpost.com/	*Welcome to WashingtonPost.com*
-.074	http://www.cbs.com/	*EYE ON THE NET @ CBS*
-.066	http://www.npr.org/	*Welcome to NPR*
-.063	http://www.telegraph.co.uk/	*Electronic Telegraph*
-.061	http://nytimesfax.com/	*TimesFax*

Figure 7: Positive and negative entries of the first non-principal eigenvector for `www.nytimes.com`.

principal eigenvectors produces distinct communities of pro-choice and pro-life pages. The issue is complicated by the existence of hub pages that link extensively to pages from both sides; but the 2nd non-principal eigenvector does produce a very clear separation. See Fig. 5.

4 Similar-Page Queries

Suppose we have found a page p that is of interest — perhaps it is an authoritative page on a topic of interest — and we want to use the link structure of the environment to discover whether there exist pages that are "similar" to p. We show how a minor modification of the framework developed above provides a novel type of link-based page similarity. It is based on the following notion. If we build an appropriate "neighborhood" T of pages around p, and p turns out to be a good authority in some community of T, then the other authorities in the same community as p will exhibit a type of linked-based similarity to p.

The algorithm is simply the following. We define the *root set* S to be k (say 200) pages that point to the initial page p. We then run the algorithm of Section 2 from this root set: we form the enlarged base set T, and find hubs and authorities in this set.

In many cases, the results can be quite compelling. For the example in Figure 6, we begin from the home page of Honda Motor Company, `www.honda.com`. We claim that it would be very difficult to automatically compile a list such as the one in this figure through text-based methods: many of the above pages consist almost entirely of images, with very little text; and the text that they do contain has very little overlap. Our approach, on the other hand, is determining, via the presence of links, what the creators of WWW pages tend to "classify" together with the given page `www.honda.com`.

In order for this method to be most effective, the initial page p should have fairly large in-degree, and "locally" be a strong authority. Otherwise, it is likely not to show up among the top authorities in the first few

communities. One variant of this phenomenon that happens quite frequently is the following. Since the home pages of search engines and computer companies have strong representation in the vicinity of essentially *every* page on the www, they often dominate the list of authorities in the principal community, regardless of the topic of the initial page p. This a case in which the communities associated with non-principal eigenvectors can be particularly valuable: it is often possible to find a strong non-principal community in which the "noise" introduced by such pages is completely eliminated, and what remains is closely related to the initial page p. This is a good example of the notion of "on-topic" versus "off-topic" communities, discussed earlier. A very clear illustration of this phenomenon is provided by a similar-page query starting from `www.nytimes.com`, the home page of the New York Times. The top authorities in the principal community consist, essentially, of a mixture of two types of pages: news organizations and computer/Internet companies. As shown in Figure 7, the first non-principal eigenvector separates this superposition sharply into its two components.

Acknowledgements. I thank Prabhakar Raghavan for invaluable on-going discussions on aspects, evaluations, and extensions of this work; Robert Kleinberg for generously sharing, as always, his insights on these problems; Rob Barrett for suggesting the use of this method on the IBM Research Intranet and providing me with the initial data; and Tryg Ager, Soumen Chakrabarti, David Gibson, Alan Hoffman, Nimrod Megiddo, Christos Papadimitriou, Sridhar Rajagopalan, and Eli Upfal for their valuable comments and suggestions.

References

[1] G.O. Arocena, A.O. Mendelzon, G.A. Mihaila, "Applications of a Web query language," *Proc. International World Wide Web Conf.*, 1997.

[2] R. Botafogo, E. Rivlin, B. Shneiderman, "Structural analysis of hypertext: Identifying hierarchies and useful metrics," *ACM Trans. Inf. Sys.*, 10(1992).

[3] J. Carrière, R. Kazman, "WebQuery: Searching and visualizing the Web through connectivity," *Proc. International World Wide Web Conf.*, 1997.

[4] Digital Equipment Corporation, *AltaVista*, `http://altavista.digital.com/`.

[5] W.E. Donath, A.J. Hoffman, "Algorithms for partitioning of graphs and computer logic based on eigenvectors of connections matrices," *IBM Technical Disclosure Bulletin*, 15(1972).

[6] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, "Indexing by latent semantic analysis," *J. American Soc. Info. Sci.*, 41(1990).

[7] M. Fielder, "Algebraic connectivity of graphs," *Czech. Math. J.*, 23(1973).

[8] M.E. Frisse, "Searching for information in a hypertext medical handbook," *Comm. ACM*, 31(1988).

[9] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, 178(1972).

[10] E. Garfield, "The impact factor," *Current Contents*, June 20, 1994.

[11] G. Golub, C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.

[12] International Business Machines, *IBM patent server*, `http://patent.womplex.ibm.com`.

[13] M.M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, 14(1963).

[14] L. Page, "PageRank: Bringing order to the Web," Stanford Digital Libraries working paper 1997-0072.

[15] P. Pirolli, J. Pitkow, R. Rao, "Silk from a sow's ear: Extracting usable structures from the Web," *Proc. ACM SIGCHI Conf. on Human Factors in Computing*, 1996.

[16] E. Rivlin, R. Botafogo, B. Shneiderman, "Navigating in hyperspace: designing a structure-based toolbox," *Comm. ACM*, 37(1994).

[17] R. Rousseau, G. Van Hooydonk, "Journal production and journal impact factors," *J. American Soc. Info. Sci.*, 47(1996).

[18] W.M. Shaw, "Subject and Citation Indexing: Part I: The clustering structure of composite representations in the cystic fibrosis document collection," *J. American Soc. Info. Sci.*, 42(1991).

[19] W.M. Shaw, "Subject and Citation Indexing. Part II: The optimal, cluster-based retrieval performance of composite representations," *J. American Soc. Info. Sci.*, 42(1991).

[20] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. American Soc. Info. Sci.*, 24(1973).

[21] E. Spertus, "ParaSite: Mining structural information on the Web," *Proc. International World Wide Web Conf.*, 1997.

[22] D. Spielman, S. Teng, "Spectral partitioning works: Planar graphs and finite-element meshes," *Proc. IEEE Symp. on Foundations of Computer Sci.*, 1996.

[23] TradeWave Corporation, *Galaxy*, `http://doradus.einet.net/galaxy.html`.

[24] World Wide Web Consortium, *World Wide Web Virtual Library*, `http://www.w3.org/vl/`.

[25] R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, D. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proc. ACM Conf. on Hypertext*, 1996.

[26] H.D. White, K.W. McCain, "Bibliometrics," in *Ann. Rev. Info. Sci. and Technology*, Elsevier, 1989.

[27] Yahoo! Corporation, *Yahoo!*, `http://www.yahoo.com`.