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. 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 for a variety of search problems on the www.

The central problem we consider is that of determining the relative "authority" of pages in such environments. This issue is central to a number of basic hypertext search tasks; for example, if the result of a query-based search consists of a large set of relevant pages, one may wish to select a small subset of the most "definitive" or "authoritative" pages to present to a user. At the same time, it is clearly difficult to formulate a definition of authority precise enough to be used in such contexts.

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.

*Preliminary versions of this paper appear in the Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998, and as IBM Research Report RJ 10076, 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.

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 [3, 13, 26, 36] and (in a context predating hypermedia) through journal citation analysis in the field of bibliometrics [37]. 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 document intuitively represents a high-quality response to a broad user query. Our algorithmic definition naturally provides an efficient means to compute this authority measure; moreover, an analysis of our method in terms of eigenvectors turns out motivate 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 [6]), 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 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"` below.

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*

The above list 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 [9] 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.

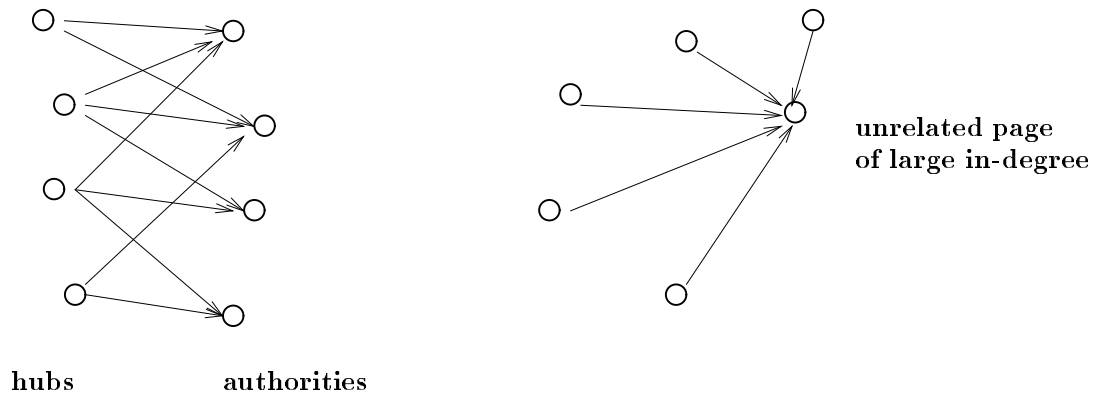

Figure 1: A dense community of hubs and authorities.

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 1; 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.

We have already discussed the motivation for returning authoritative pages in response to a search. Hub pages are also valuable, since they typically provide a well-situated starting point from which a user can gather many relevant, authoritative pages on the query topic. In the context of the WWW, these pages often arise as hand-assembled collections of links to resources on topics of interest to the creator of the page; the distinctive properties of such pages have been observed in a number of sources. For example, Duffy and Yacovissi write

> In many respects, the mere content on the Web is less significant than the way in which the network fosters this simultaneous sense of serendipitous discovery and "connectedness". Consider those ubiquitous hyperlinked lists of favorite sites that every individual home page impresario apparently feels compelled to assemble. And look beyond the grass-roots Webheads too, if you want. You may be surprised at how often similar – though usually more carefully constituted – jump-lists turn up at the Web sites of corporations as well [8].

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"` [5]); 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. [7, 11, 34]); 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. We will be presenting a number of such examples in the following sections. 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[38], Galaxy [14], Zia [39], and the distributed WWW Virtual Library [35]. These hierarchies 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. In Section 5.1 we compare the results of our technique to the contents of these and other hierarchies, for ten topics from YAHOO's *Health/Medicine* directory. Third, we are interested in evaluating the extent to which the communities discovered by our method have a *robust* identity — that is, the extent to which several different root sets on the same "topic" produce similar communities of hubs and authorities. In Section 5.2, we discuss several experiments along these lines, including a comparison of root sets produced by issuing the same query to a number of different search engines, and a comparison of root sets produced by issuing a query term in a number of different language.

Finally, in Section 6, we consider the problem of determining how "broad" a query topic must be in order for our method to produce reliable sets of hubs and authorities. In particular, we consider 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 Section 6 that non-principal eigenvectors can be a very effective way to produce relevant results even in the face of this phenomenon. Specifically, 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, since we need only approximate the coordinates of large absolute value in the eigenvectors we use, we find that reliable results can be obtained by using a simple iterative algorithm with a very small number of iterations. Finally, as we have already discussed, our method produces meaningful output using only a small fragment of the WWW in answering 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, has been 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 [19], 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* [37] — the study of written documents and their citation structure. Some related work has also been done in the hypertext research community. This work has focused predominantly on the use of citations and/or explicit hyperlinks as a means of clustering and enhancing relevance judgments.

Two basic measures of document similarity to emerge from the study of bibliometrics are *bibliographic coupling* [20] and *co-citation* [32]. For two documents p and q, the former quantity is equal to the number of documents cited by both p and q, and the latter quantity is the number of documents that cite both p and q. We will see that these two quantities arise inside the analysis of our method. Shaw [30, 31] uses a combination of these measures, together with some textual measures, as part of a graph-based clustering algorithm. Documents constitute the nodes of an undirected graph whose edge weights are measures of similarity; edges are deleted in order of increasing weight, producing a hierarchical clustering. (See e.g. [25].) Schwanke and Platoff [29] discuss an interesting application of these bibliometric measures for clustering purposes in a completely different realm — that of analyzing the relationships among modules in a large software system.

There has also been work in bibliometrics on using citation counts to assess the "impact" of scientific journals; this is more closely related to the issue we are considering here. The classic work in this area is that of Garfield [15]; see also e.g. [16, 27]. In addition to basic differences in the algorithms themselves, a fundamental difference between our method and the classical bibliometric work lies in our explicit separation of *hubs* and *authorities*, and investigation of the equilibrium that exists between them — as discussed above, this is a phenomenon that perhaps makes more sense in the context of hypermedia such as the WWW, where one finds several categories of participants, than it does in the classic bibliometric setting of scientific journals, which ostensibly serve a uniform role. At the same time, it has been observed in bibliometric studies that *review journals* often constitute exceptions to certain basic principles [16, 27]; it would be interesting to see whether the distinction between hubs and authorities would be useful in this regard.

There has been some work on using hyperlinks for clustering and searching in the hypertext research community. Rivlin, Botafogo, and Schneiderman [3, 26] use basic graph-theoretic notions such as connectivity, as well as "compactness" measures based on node-to-node distances, to identify clusters in the graph of a hypertext environment. Weiss et al. [36] define similarity measures among

pages in a hypertext environment based on the link structure; these measures are generalizations of *co-citation* and *bibliographic coupling* to allow for arbitrarily long chains of references. Larson [22] performs a co-citation analysis of a set of pages relevant to a sample query and generates clusters by dimension-reduction techniques. Finally, Frisse [13] describes a method that is applicable in a tree-structured environment: the *relevance* of a page with respect to a query is also based on the relevance of its descendants in the tree.

More recently, Pirolli, Pitkow, and Rao [24] 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 [33] have described frameworks for constructing WWW queries from a combination of term-matching and link-based predicates. Page [23] 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 [4] propose a link-based method for visualizing and ranking the results of queries returned by WWW search engines. Their method is essentially to (1) enlarge the set S of returned pages to include any page joined to a member of S via a link (in either direction); and then (2) rank each page p in this enlarged set according to the number of pages connected to p via links (again, in either direction). Although the notion of augmenting search engine results to a "one-step neighborhood" is a basic step in our method, the algorithmic component of our work differs significantly from that of [4]. In particular, we make crucial use of the directionality of hyperlinks, including the explicit distinction of *hubs* and *authorities*; our ranking of pages is not obtained by a direct counting of neighbors in the link structure; and our framework naturally allows for the construction of multiple communities of ranked pages.

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 the 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

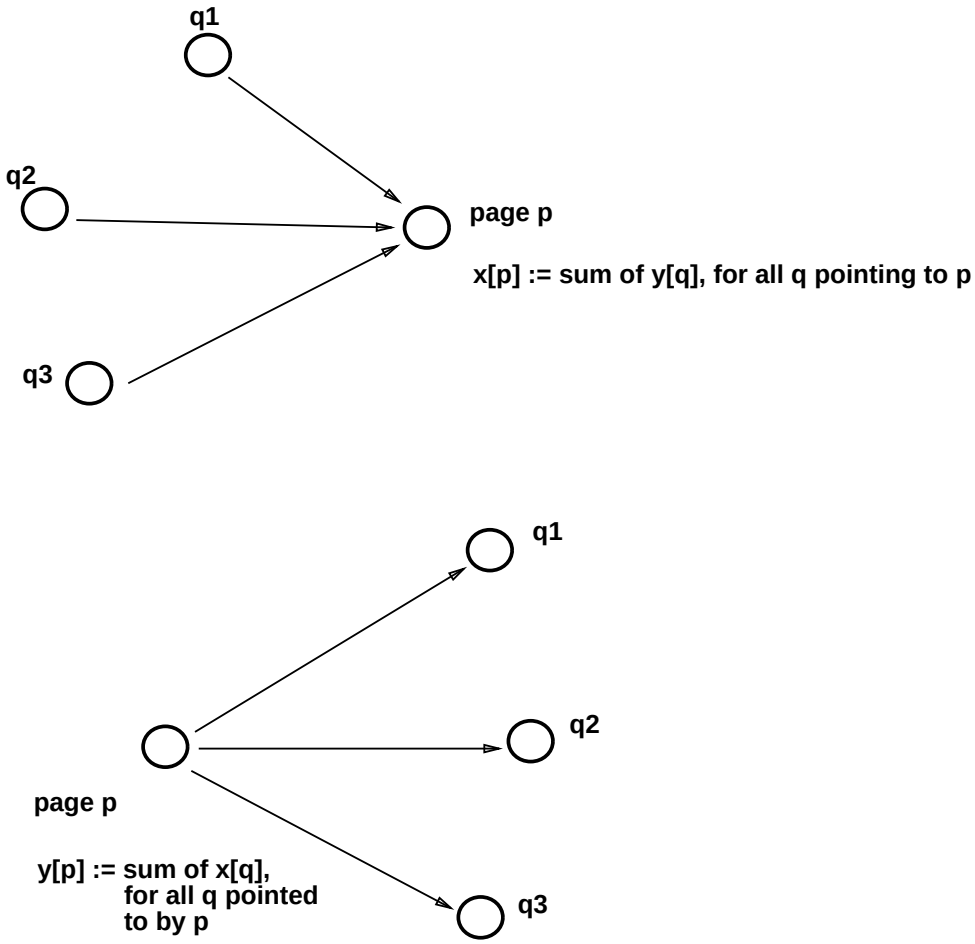

Figure 2: The basic operations.

$\mathcal{I}$ 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. (See Figure 2.)

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 $(1, 1, 1, \ldots, 1)$.
Initially set $x \leftarrow z$; $y \leftarrow z$.
For $i = 1, 2, 3, \ldots$

```
    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*)
    (see Theorem 2.1).
Return (x*, y*) as the authority and 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 $(\mathcal{O}\mathcal{I})(x, y)$ denote the result of applying the $\mathcal{I}$ operation followed by the $\mathcal{O}$ operation (and normalizing the vectors obtained). Let $(\mathcal{O}\mathcal{I})^n(x, y)$ 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 $(x_n, y_n) = (\mathcal{O}\mathcal{I})^n(z, z)$.

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:

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

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. [17]) 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 ($\dagger$)), 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 ($\dagger$).) x^* is the principal eigenvector of $A^T A$, and y^* is the principal eigenvector of AA^T.*

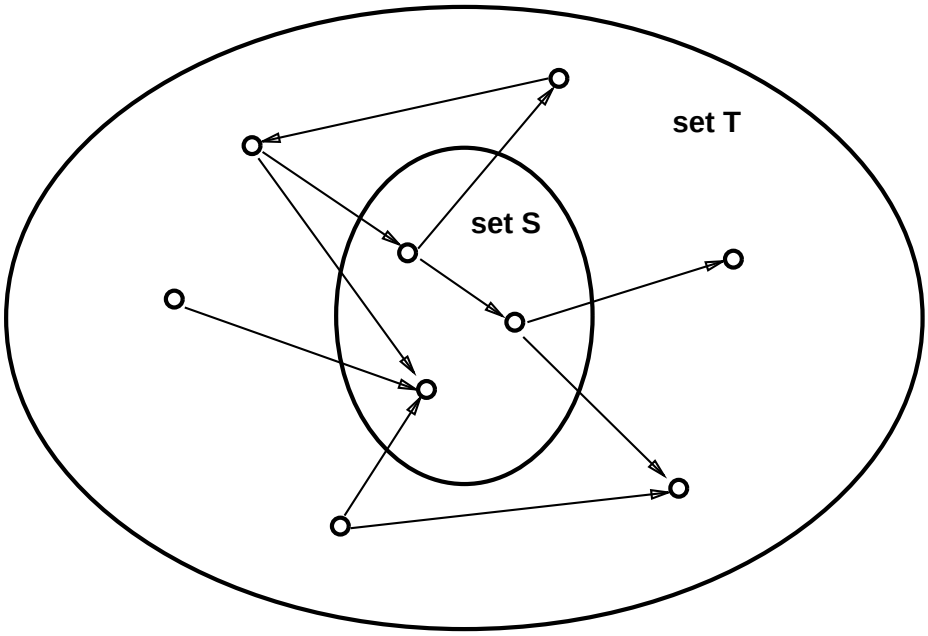

Figure 3: Expanding the root set into a base set.

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 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.

It is interesting to note that the $(i, j)^{\text{th}}$ entry of $A^T A$ gives the number of pages that point to both p_i and p_j; the $(i, j)^{\text{th}}$ entry of AA^T gives the number of pages pointed to by both p_i and p_j. Thus, these individual matrix entries correspond to the notions of *co-citation* and *bibliographic coupling* discussed above.

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_0 to be the (directed) subgraph of the WWW induced on the set T. We now distinguish between two types of WWW links. We say that a link is *transverse* if it is between pages with different domain names, and *intrinsic* if it is between pages with the same domain name. By "domain name," we mean here the first level in the URL string associated with a page. Since intrinsic links are very often created simply to help users navigate the infrastructure of a site, they tend to convey much less information than transverse links about the authority of the pages they point to. Thus, we delete all intrinsic links from the graph G_0, keeping only the edges corresponding to transverse links; this results in a graph G. (Clearly there are a variety of more sophisticated ways of handling the information provided by the URL hierarchy; this remains an interesting issue.)

(4) Finally, we run the iterative algorithm described above 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

("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*

Above, we list the top six authorities found by the algorithm. Each line 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)`, we obtain the following authorities. (For the second query, the syntax indicates that both words must appear in the initial pages returned by the search engine.)

(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	

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 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. We will return to this issue in Section 6, and study some cases in which the principal community was not so closely tied to the initial query topic.

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, it is possible that there will 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)` (an example drawn from [5]).

(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 begin by noting 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 $\omega_i(AA^T) = A\omega_i(A^TA)$.*

Proof. Let $\lambda_i = \lambda_i(A^T A)$ be an eigenvalue of $A^T A$, and $v_i = \omega_i(A^T A)$ the associated eigenvector. It suffices to show that Av_i is an eigenvector of AA^T, and that the associated eigenvalue is λ_i. We observe

$$(AA^T)(Av_i) = A((A^T A)v_i) = A(\lambda_i v_i) = \lambda_i(Av_i).$$

■

Thus, each pair of eigenvectors $x_i^* = \omega_i(A^T A)$, $y_i^* = \omega_i(AA^T)$, 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* [7, 11, 34], 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.

Basic Results

We now give some examples of the types of communities one obtains by this methods, using the queries mentioned above.

One interesting phenomenon that arises is the following. The pages with large coordinates in the first few non-principal eigenvectors tend to recur, so that essentially the same community of hubs and authorities will often be generated by several of the strongest non-principal eigenvectors. Of course, despite being similar in their large coordinates, these eigenvectors remain orthogonal due to differences in the coordinates of smaller absolute value. As a result, one obtains fewer distinct communities than might otherwise be expected from a set of non-principal eigenvectors.

This notion is also reflected in the output below, where we have selected (by hand) several distinct communities from among the first few non-principal eigenvectors. The identification of "distinct" communities might be an interesting task to make automatic, though this does not appear to be a particularly difficult challenge.

We issue the first query as `(jaguar jaguars)`, simply as one way to search for either the word or its plural. For this query, the strongest communities concerned the Atari Jaguar product, the NFL football team from Jacksonville, and the automobile.

(jaguar jaguars) Authorities: principal eigenvector

.370	http://www2.ecst.csuchico.edu/~jschlich/Jaguar/jaguar.html	
.347	http://www-und.ida.liu.se/~t94patsa/jserver.html	
.292	http://tangram.informatik.uni-kl.de:8001/~rgehm/jaguar.html	
.287	http://www.mcc.ac.uk/ dlms/Consoles/jaguar.html	*Jaguar Page*

(jaguar jaguars) Authorities: 2nd non-principal vector, positive end

.255	http://www.jaguarsnfl.com/	*Official Jacksonville Jaguars NFL Website*
.137	http://www.nando.net/SportServer/football/nfl/jax.html	*Jacksonville Jaguars Home Page*
.133	http://www.ao.net/~brett/jaguar/index.html	*Brett's Jaguar Page*
.110	http://www.usatoday.com/sports/football/sfn/sfn30.htm	*Jacksonville Jaguars*

(jaguar jaguars) Authorities: 3rd non-principal vector, positive end

.227	http://www.jaguarvehicles.com/	*Jaguar Cars Global Home Page*
.227	http://www.collection.co.uk/	*The Jaguar Collection - Official Web site*
.211	http://www.moran.com/sterling/sterling.html	
.211	http://www.coys.co.uk/	

For the query `("randomized algorithms")`, none of the strongest communities could be said to be precisely on the query topic, though they all consisted of thematically related pages on a closely related topic. They included home pages of theoretical computer scientists, compendia of mathematical software, and pages on wavelets.

("randomized algorithms") Authorities: 1st non-principal vector, positive end

.125	http://theory.lcs.mit.edu/~goemans/	*Michel X. Goemans*
.122	http://theory.lcs.mit.edu/~spielman/	*Dan Spielman's Homepage*
.122	http://www.nada.kth.se/~johanh/	*Johan Hastad*
.122	http://theory.lcs.mit.edu/~rivest/	*Ronald L. Rivest : HomePage*

("randomized algorithms") Authorities 1st non-principal vector, negative end

-.00116	http://lib.stat.cmu.edu/	*StatLib Index*
-.00115	http://www.geo.fmi.fi/prog/tela.html	*Tela*
-.00107	http://gams.nist.gov/	*GAMS : Guide to Available Mathematical Software*
-.00107	http://www.netlib.org	*Netlib*

("randomized algorithms") Authorities 4th non-principal vector, negative end

-.176	http://www.amara.com/current/wavelet.html	*Amara's Wavelet Page*
-.172	http://www-ocean.tamu.edu/~baum/wavelets.html	*Wavelet sources*
-.161	http://www.mathsoft.com/wavelets.html	*Wavelet Resources*
-.143	http://www.mat.sbg.ac.at/~uhl/wav.html	*Wavelets*

We mentioned that the two communities associated with the positive and negative ends of the same non-principal eigenvector are often "well-separated" in the link structure. 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-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 in fact the 2nd non-principal eigenvector produces a very clear separation:

(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/	
-.144 http://www.clark.net/pub/jeffd/factbot.html	*The Right Side of the Web*
-.144 http://www.catholic.net/HyperNews/get/abortion.html	

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 first 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 starting 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. In the following examples, we begin from the home page of Honda Motor Company, `www.honda.com`, and the New York Stock Exchange, `www.nyse.com`.

(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*

(www.nyse.com) Authorities: principal eigenvector

.208 http://www.amex.com/	*The American Stock Exchange - The Smarter Place to Be*
.146 http://www.nyse.com/	*New York Stock Exchange Home Page*
.134 http://www.liffe.com/	*Welcome to LIFFE*
.129 http://www.cme.com/	*Futures and Options at the Chicago Mercantile Exchange*
.120 http://update.wsj.com/	*The Wall Street Journal Interactive Edition*
.118 http://www.nasdaq.com/	*The Nasdaq Stock Market Home Page - Reload Often*
.117 http://www.cboe.com/	*CBOE - The ChicagoBoard Options Exchange*
.116 http://www.quote.com/	*1- Quote.com - Stock Quotes, Business News, Financial Market*
.113 http://networth.galt.com/	*NETworth*
.109 http://www.lombard.com/	*Lombard Home Page*

Note the difficulties inherent in compiling such lists 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 pages `www.honda.com` and `www.nyse.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. However, even when p does not show up among the top authorities, the resulting output is often valuable as a a type of "broad-topic classification" of p. We illustrate this with the home page of the ACM Special Interest Group on Algorithms and Computation Theory, `sigact.acm.org`. The page itself did not rank highly in any of the first few eigenvectors; however, the principal community is quite revealing as a summary of the strongest authorities in the "vicinity" of SIGACT.

(sigact.acm.org) Authorities: principal eigenvector

.197 http://www.siam.org/	*Society for Industrial and Applied Mathematics*
.166 http://dimacs.rutgers.edu/	*Center for Discrete Mathematics and Theoretical Computer Science*
.150 http://www.computer.org/	*IEEE Computer Society*
.148 http://www.yahoo.com/	*Yahoo!*
.145 http://e-math.ams.org/	*e-MATH Home Page*
.141 http://www.ieee.org/	*IEEE Home Page*
.140 http://glimpse.cs.arizona.edu:1994/bib/	*Computer Science Bibliography Glimpse Server*
.129 http://www.eccc.uni-trier.de/eccc/	*ECCC - The Electronic Colloquium on Computational Complexity*
.129 http://www.cs.indiana.edu/cstr/search	*UCSTRI — Cover Page*
.118 http://euclid.math.fsu.edu/Science/math.html	*The World-Wide Web Virtual Library: Mathematics*

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. First, consider the authorities in the principal community.

(www.nytimes.com) Authorities: principal eigenvector

.287	http://www.yahoo.com/	*Yahoo!*
.181	http://www.nytimes.com/	*The New York Times on the Web*
.170	http://www.usatoday.com/	*USA TODAY*
.165	http://www.cnn.com/	*CNN Interactive*
.124	http://www.mckinley.com/	*Welcome to Magellan!*
.120	http://www.altavista.digital.com/	*AltaVista Search: Main Page*
.119	http://www.excite.com/	*Excite*
.117	http://www.microsoft.com/	*Welcome to Microsoft*
.108	http://www.whitehouse.gov/	*Welcome to the White House*
.107	http://www.lycos.com/	*Lycos, Inc. Home Page*

The above list consists, essentially, of a mixture of two types of pages: news organizations and computer/Internet companies. As shown below, the first non-principal eigenvector separates this superposition into its two components: it has computer companies at its positive end and news organizations at its negative end.

(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*

5 Searchable Hierarchies and Other Experiments

We have now seen the application of the basic framework for discovering authoritative pages on a query topic, finding multiple communities of thematically related pages, and finding pages similar to an initial query page. In some respects, it is easier to verify the effectiveness of the algorithm at the latter two tasks than at the first — one can confirm, for example, that the pages listed together with `www.honda.com` are all automobile manufacturers; or that the two communities listed for the query `(abortion)` represent different sides of the issue. The notion of *authority*, however, remains somewhat more elusive; although one can attempt to evaluate whether the pages returned seem to be "authoritative," can one find a way to allow for a more concrete evaluation of the algorithm?

We claim that the *searchable hierarchies* available on the WWW provide us with one approach towards achieving this; this will be the topic of Section 5.1. We mentioned examples of such searchable hierarchies in the introduction (e.g. YAHOO[38], Galaxy [14], Zia [39], and the WWW Virtual Library [35]); they are lists of authoritative pages, compiled by humans, on a variety of broad search topics. Such hierarchies provide us both with externally generated lists of query topics, and with lists of authoritative pages on these topics against which to compare our results.

In Section 5.2, we take up a different issue: we show some respects in which different methods of producing a root set for the same query topic lead to very similar communities of hubs and authorities.

5.1 Searchable Hierarchies

We have tested the algorithm on the topics in the YAHOO directories *Health/Medicine*, *Science/Physics*, and *Entertainment/Movies/Genres*, as well as portions of several others. To illustrate the comparisons one can make between the output of our method and the contents of WWW searchable hierarchies, we consider ten topics drawn from the YAHOO *Health/Medicine* directory: acupuncture, anatomy, anesthesiology, audiology, cardiology, dermatology, endocrinology, epidemiology, gastroenterology, and hematology. For the sake of concreteness, we consider only the top 20 hubs and authorities in the principal community. (See Figure 4.)

Although the queries were drawn explicitly from YAHOO, pages from several different searchable hierarchies appeared as high-scoring hub pages for many of the topics. These included both general–purpose hierarchies which attempt to represent all topics, and specialized medical hierarchies which contain pages only for a range of medical key words. The former category contains sites such as those mentioned above; the latter category contains sites such as the MedMark index at `medmark.bit.co.kr`, and the MedWeb index at `www.gen.emory.edu/medweb`.

The table in Figure 4 provides the following information: from among the top 20 hubs and top 20 authorities in the principal community for each query term, it lists the set of pages referenced by each of YAHOO, Galaxy, and Zia. Referenced pages are labeled by their rank, with the prefix 'A' or 'H'; a dash indicates that the hierarchy did not contain a page for the associated topic. We make the following observations about the experiment.

• Of the three hierarchies, the relevant page from YAHOO itself was among the top 20 hub pages once (as H9 under (audiology)), the relevant Zia page was among the top 20 hub pages once (as H10 under (audiology)), and the relevant Galaxy page was among the top 20 hub pages twice (as

query	YAHOO	Zia	Galaxy
acupuncture	A3, A4, A6	A3, A6	— —
anatomy	A1, A2, A4, A18	— —	— —
anesthesiology	A4	A4, A7, H0	— —
audiology	A0, A1, A3, A7, A14	A0, A1, A3, A7, A14	— —
cardiology	A16	A16	A3, A5, A9, A10, A12
dermatology			A4
endocrinology			A10, A14
epidemiology	H5	H5	— —
gastroenterology	A0, A2	A2, A3	A3, H9, H12
hematology	A0	A0	A0, A1, A2, A5, A7, A11, A12, A13, A15, A17, A18

Figure 4: Hubs/authorities referenced by 3 searchable hierarchies, on topics from *Health/Medicine*.

H14 under (gastroenterology) and H10 under (hematology)).

• The identity of the top hub page (H0) can be summarized as follows. The WWW Virtual Library hierarchy contained pages for two of the topics — anesthesiology and epidemiology. For both of these topics, it emerged as the top hub page. For five of the queries, the top hub page was the relevant page from the MedMark hierarchy at `medmark.bit.co.kr`. For the remaining three topics, the top hub page contained pointers to many pages relevant to the query, but did not appear to belong to a larger searchable hierarchy. Thus the top hub pages (and authorities) were relevant to the initial query for all ten topics. This can clearly be attributed in some measure to the quantity of professionally assembled hub pages for these topics.

• The general-purposes hierarchies YAHOO, Galaxy, and Zia did not score as highly as some of the focused medical hierarchies, though they consistently referenced authorities in the top 20. The WWW Virtual Library, while also a general purpose hierarchy, scored extremely well.

5.2 Alternative Root Sets

For broad-topic queries, our method produces meaningful results despite starting from a very small sample of relevant pages in the initial root set. This implies that the communities of hubs and authorities that are produced exhibit a certain type of "robustness." Here, we give some additional indications of the way in which the communities resulting from our search method are insensitive to the precise choice of pages in the initial root set.

Multiple Search Engines One straightforward way to produce different root sets for the same query string is to issue the query to several different term-based search engines, such as AltaVista [6], Infoseek [18], and Excite [10]. Typically, issuing the query to several search engines will have the effect of producing root sets that have very little intersection with one another. We have run

a number of tests of this form; we have found that the main communities tend to have similar structure as one varies the root set, although the eigenvectors with which they are associated can change. Thus, for example, the hubs and authorities associated with the principal eigenvector under one root set S can become associated with a non-principal eigenvector under a different root set S'.

As an illustration, we consider the query `("web browsers")`. in Section 2, we showed the authorities in the principal community for the root set provided by AltaVista; here we show that starting from Infoseek or Excite produces similar communities. (In case of Excite, the *principal* community is quite different, but one finds a community of browser manufacturers associated with the third non-principal eigenvector.)

("web browsers") (via Infoseek) Authorities: principal eigenvector		
.227	http://www.ncsa.uiuc.edu/SDG/Software/WinMosaic/HomePage.html	*NCSA Windows Mosaic Home Page*
.197	http://www.interport.net/slipknot/slipknot.html	*.... SlipKnot Home Page ....*
.195	http://galaxy.einet.net/EINet/WinWeb/WinWebHome.html	*winWeb and MacWeb*
.192	http://www.microsoft.com/	*Welcome to Microsoft*
.185	http://home.mcom.com/home/welcome.html	*Welcome to Netscape*
.179	http://www.ncsa.uiuc.edu/General/NCSAHome.html	*Welcome to NCSA*

("web browsers") (via Excite) Authorities: 3rd non-principal vector, negative end		
-.560	http://home.netscape.com/	*Welcome to Netscape*
-.252	http://www.microsoft.com/	*Welcome to Microsoft*
-.231	http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html	*NCSA Windows Mosaic Home Page*
-.224	http://www.microsoft.com/ie/	*Microsoft Internet Explorer*
-.215	http://galaxy.einet.net/EINet/MacWeb/MacWebHome.html	*winWeb and MacWeb*
-.189	http://www.ncsa.uiuc.edu/SDG/Software/MacMosaic/MacMosaicHome.html	*Mosaic for the Macintosh*

Cross-Linguistic Queries Another method to obtain multiple root sets for the same query topic is to issue the query to a search engine in several different languages. This experiment is additionally interesting in that it emphasizes another advantage of the link-based nature of our method; since it does not make use of the text of pages, it can discover good hubs and authorities in multiple languages, even grouping them together in a common community provided that they link densely enough to one another. We have observed this phenomenon in a variety of cases: the method often finds relevant hubs and authorities written in languages other than English, even for queries issued in English.

We issued a number of the topics in YAHOO's *Science/Physics* list in English, French, and German. We have found that the principal community for the English version often appears in a relatively similar form as a non-principal community for the French and German versions. As an example, we consider the first of the *Science/Physics* topics:

(astrophysics) Authorities: principal eigenvector		
.226	http://fits.cv.nrao.edu/www/astronomy.html	*AstroWeb: Astronomy/Astrophysics on the Internet*
.189	http://cdsweb.u-strasbg.fr/Simbad.html	*The SIMBAD astronomical database*
.189	http://www.aas.org/	*American Astronomical Society Home Page*
.183	http://heasarc.gsfc.nasa.gov/	*HEASARC/GSFC Home Page*
.175	http://adsabs.harvard.edu/abstract-service.html	

.169 http://cdsweb.u-strasbg.fr/CDS.html *CDS, Strasbourg*
.161 http://adswww.harvard.edu/ *The NASA Astrophysics Data System Home Page*

(astrophysique) Authorities: 8^{th} non-principal vector, negative end
-.253 http://cdsweb.u-strasbg.fr/CDS.html *CDS, Strasbourg*
-.239 http://adswww.harvard.edu/ *The NASA Astrophysics Data System Home Page*
-.208 http://cdsweb.u-strasbg.fr/Simbad.html *The SIMBAD astronomical database*
-.189 http://adsabs.harvard.edu/abstract-service.html
-.153 http://fits.cv.nrao.edu/www/astronomy.html *AstroWeb: Astronomy/Astrophysics on the Internet*
-.134 http://www.hq.eso.org/eso-homepage.html *European Southern Observatory Homepage*
-.122 http://info.er.usgs.gov/network/science/astronomy/index.html *Astronomy and Space Science*

(astrophysik) Authorities: 7^{th} non-principal vector, negative end
-.306 http://adswww.harvard.edu/ *The NASA Astrophysics Data System Home Page*
-.273 http://cdsweb.u-strasbg.fr/Simbad.html *The SIMBAD astronomical database*
-.273 http://adsabs.harvard.edu/abstract-service.html
-.237 http://aibn55.astro.uni-bonn.de:8000/
-.186 http://www.univ-rennes1.fr/ASTRO/astro.english.html *Astronomical pictures & animations*
-.173 http://aorta.tat.physik.uni-tuebingen.de/
-.139 http://fits.cv.nrao.edu/www/astronomy.html *AstroWeb: Astronomy/Astrophysics on the Internet*

6 Diffusion and Lexical Scores

Finally, we turn to a discussion of some of the typical ways in which the algorithm fails, and some possible extensions of the method to make it more robust against the most common difficulties. We will be focusing on the query-based experiments of Sections 2 and 3, in which a root set S is generated by a search engine, a set T is grown around it, and hubs and authorities are determined for the set T. The basic phenomenon we investigate here is what one could call *diffusion*: the algorithm converges to a set of hubs and authorities that are not focused on the original topic. A large fraction of the cases in which diffusion occurs share several basic features: the query topic is relatively specific, and there is a "generalization" of the query topic with much greater representation on the WWW. In these cases, the principal community of hub and authorities is often relevant to this generalization of the initial topic.

Basic Examples

Some simple examples will make this notion of diffusion more concrete. First, consider the query `("medical conferences")` from the YAHOO *Health/Medicine* list. Although AltaVista indexes roughly 600 pages containing the term, the algorithm essentially converges to authoritative pages for the topic "medicine" in general:

("medical conferences") Authorities: principal eigenvector
.087 http://www.cdc.gov/ *Centers for Disease Control and Prevention Home Page*
.083 http://www.ohsu.edu/cliniweb/ *Cliniweb*
.081 http://www.bmj.com/bmj/ *BMJ*

It is not surprising that this should have happened: pages on medical conferences typically link to a large number of general medical resources, and hence these acquire a lot of authority. More interesting, perhaps, is to ask why a similar phenomenon did not occur for the ten medical topics considered in Section 5.1. Although the issue is clearly quite subtle, one can observe that (a) the ten earlier topics are considerably "larger" than the current one, in that AltaVista indexes roughly $20,000$ pages for most of them; and (b) there are correspondingly more resource pages focused on these topics, which can "freeze" the authority weight at more specific pages and prevent it from diffusing to more general ones.

A strictly analogous phenomenon occurs for the query `("WWW conferences")`. AltaVista indexes roughly 300 pages containing the term; but unfortunately, it is a specialization of the largest WWW topic of all: the Web itself.

("WWW conferences") Authorities: principal eigenvector
.088 http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-new.html *The What's New Archive*
.088 http://www.w3.org/hypertext/DataSources/WWW/Servers.html *World-Wide Web Servers: Summary*
.087 http://www.w3.org/hypertext/DataSources/bySubject/Overview.html *The World-Wide Web Virtual Library*

These two examples demonstrate what appears to be by far the most common form of *diffusion*: the authorities associated with the principal eigenvector correspond to a generalization of the initial query topic. To reiterate the basic reason for this phenomenon in a slightly different way: Once the set T of pages has been constructed, the query string is ignored, and hence the primary hubs and authorities produced will simply be consistent with whatever topic best "fits" the set of pages in T.

A Lexical Scoring Function

There is a range of techniques one could try implementing to help prevent diffusion; unfortunately, many of them strongly interfere with the positive features of the algorithm as it stands. Specifically, it is tempting to weight each page by a lexical score derived from the query string, and incorporate these weights into the iterations of the algorithm as it computes hubs and authorities. However, while this would undoubtedly keep the underlying topic more strongly in focus, there are many respects in which it would hurt the algorithm's ability to perform the basic task of locating authoritative sources. We have discussed this already in the introduction, where we noted that many of the authoritative pages one would like to find do not contain the query string. (Rather, the set of pages that do contain the string also link densely to the desired authorities.)

We propose the following alternative approach, which allows the communities to develop as before, and only re-introduces the query term once they have been constructed. It takes advantage of a point that has been brought up earlier: by producing multiple communities of hubs and authorities, one can try to distinguish the set of "on-topic" communities from the set of "off-topic" ones. Thus, we first run the basic algorithm on the base set T, producing a number of different communities. (We noted earlier that only a relatively small number of communities are associated with eigenvalues of non-trivial magnitude, and we clearly wish to restrict ourselves to these.) We then apply a lexical scoring function to each community as a whole, and rank the communities

according to this function. Of course, any attempt to use a lexical scoring function will suffer to a greater or lesser degree from the problem discussed above, that many authorities do not explicitly use the initial query term. However, by computing a single total score for all the pages in one community, one can partially offset this effect: provided only that *some* number of high-scoring pages in the relevant community use the term sufficiently frequently, one hopes that the overall score will be relatively large.

There are many options for the scoring function to use. We report here on a set of experiments performed using the following extremely simple one:

(i) For each of the communities C being considered, choose the top five hubs and top five authorities to form a set R_C of *representative pages*.

(ii) For a page p, and a query string s, let $\alpha_s(p)$ denote the number of times the string s occurs in the page p.

(iii) The score for the community C is then

$$\alpha_s(C) = \sum_{p \in R_C} \alpha_s(p).$$

It is worth commenting on the most basic variations that are possible. First of all, rather than arbitrarily choosing a representative set for the community C, one could consider the authority/hub weights (x_i^*, y_i^*) that define C and compute a cumulative weighted score:

$$\tilde{\alpha}_s(C) = \sum_p |x_i^*(p)| \alpha_s(p) + \sum_p |y_i^*(p)| \alpha_s(p).$$

Although this is perhaps aesthetically cleaner, we favored our method for two main reasons. First, our score can be computed without maintaining knowledge of all the weights associated with each community. Second, each community will be represented to the user as a small representative set; and hence if one is scoring the communities so as to improve the order in which they are presented, there is an argument for computing the score based only on what the user will actually see.

Defining the function α_s which simply *counts* the number of occurrences of the term s is undoubtedly too crude. At another extreme, one could define $\beta_s(p)$ to be 1 if the string s appears in p, and 0 otherwise, and use this to score communities. We ran all the experiments in this section using β_s in place of α_s, and achieved qualitatively similar results. In the context of this scoring framework, the best function to use is most likely something that is not based purely on term-matching; for example, one could use a scoring function derived from a technique such as *latent semantic indexing* [9], although we have not tested this here.

Despite the crudeness of our scoring function, it has proved to be surprisingly effective in many of our tests. The clustering performed by our algorithm may provide one reason why pure term-counting is more effective in this setting than in others. As we have discussed earlier, communities tend to exhibit strong uniformity in content, since they are formed on the basis of link density. Thus we are able to use the term frequency in a *set* of related pages, rather than the less reliable notion of term frequency in a *single* page. We also appear to be helped by certain properties of hub pages; we have observed that hub pages for a given topic tend to be very rich in the individual terms associated with that topic.

As an example, we consider using the function α_s to rank the communities associated with the first 20 eigenvectors for the two queries discussed at the beginning of this section. For `("medical conferences")`, the highest-scoring among these communities was associated with the 6^{th} non-principal eigenvector; as its top hub page, it produced the *Medical Conferences* page from the MedWeb searchable hierarchy. For `("WWW conferences")`, the highest-scoring community among the top 20 was associated with the 11^{th} eigenvector:

("WWW conferences") Authorities: 11^{th} non-principal vector, negative end

-.097 http://www.igd.fhg.de/www95.html	*Third International World-Wide Web Conference*
-.091 http://www.csu.edu.au/special/conference/WWWWW.html	*AUUG'95 and Asia-Pacific WWW'95 Conference*
-.090 http://www.ncsa.uiuc.edu/SDG/IT94/IT94Info.html	*The Second International WWW Conference '94*
-.083 http://www.w3.org/hypertext/Conferences/WWW4/	*Fourth International World Wide Web Conference*
-.079 http://www.igd.fhg.de/www/www95/papers/	*WWW'95: Papers*

Term Mixtures

When a query is composed of more than one term, the above method produces a lexical score for *each* term. Thus, when there are $m > 1$ terms in the query, the resulting score is a vector with m coordinates, and we face the problem that a set of such vectors has no natural total ordering. Here we discuss some natural approaches to ranking communities in this situation.

Before discussing candidates for total orders, it is worth mentioning the natural partial order: if a_1 and a_2 are m-coordinate vectors, we write $a_1 \preceq a_2$ if each coordinate of a_1 is less than or equal to the corresponding coordinate of a_2. Among the set of score vectors produced, we will say that the vector a is *maximal* if there is no $a' \neq a$ such that $a \preceq a'$. Presumably one wants to arrange things so that the highest-scoring community is a maximal one; and in most of our experiments, the set of maximal communities has proved to be considerably smaller than the full set of communities under consideration.

In order to make these notions more concrete, let us consider the following example from the YAHOO *Entertainment/Movies* list: the query `(+movies +awards)`. In a fashion similar to the previous examples, the primary community diffused to the more general topic "movies":

(+movies +awards) Authorities: principal eigenvector

.291 http://www.disney.com/	*Disney.com Home Page - Welcome*
.278 http://www.hollywood.com/	*Hollywood Online*
.217 http://www.paramount.com/	*Paramount Pictures Online*

In fact it turned out, by inspection, that none of the communities associated with the first 20 eigenvectors were highly relevant to the query; and so we also performed tests on the set of communities associated with the first 50 eigenvectors.

For each of the 99 communities examined, our term-based scoring method produced a two-coordinate vector (one coordinate for each of "movies" and "awards"). In Figure 5, we provide a scatter-plot of the resulting 99 points in the plane. Seven of the 99 points are maximal. It is also interesting to consider the *positive hull* of the point set: by this we mean the set of all points that

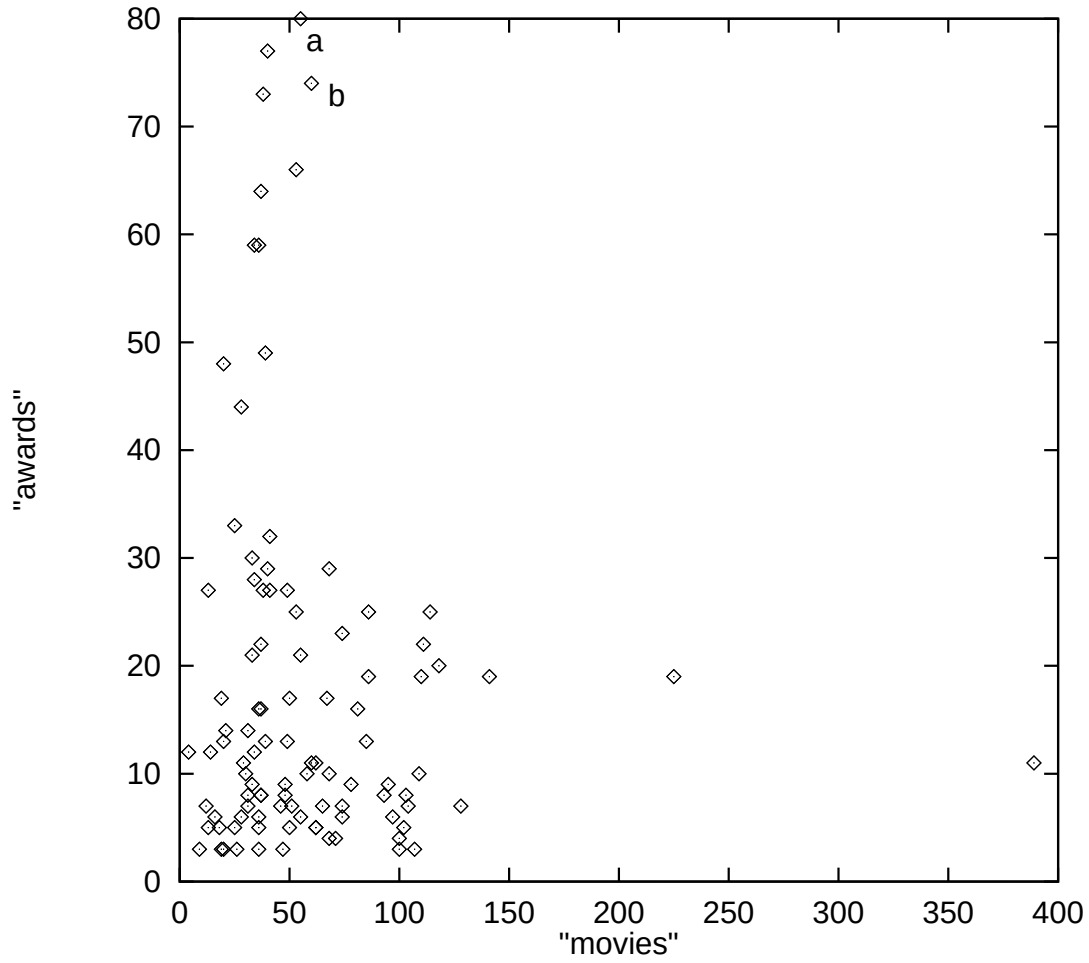

Figure 5: Term occurrences in `(+movies +awards)`

lie on a line of negative slope which does not separate the point set. Only two of the points lie on the positive hull.

The two maximal points labeled "a" and "b" in Figure 5 turned out to be arguably the most relevant communities. It is also interesting to note that they were associated with opposite ends of the same non-principal eigenvector; this corresponds, to some extent, to the heuristic intuition behind spectral graph partitioning, as discussed in Section 3. For the community labeled "a", the top five hub pages include the YAHOO and Zia pages for *Movies/Academy Awards*; the top five authorities are as follows:

(+movies +awards) Authorities: 37^{th} non-principal vector, negative end

-.118 http://www.ampas.org/	*The Academy of Motion Picture Arts and Sciences*
-.110 http://www.mnet.fr/dian.ying/	*Index dian ying*
-.108 http://ddv.com/Oscarnet/	
-.108 http://oscars.guide.com/	*THE ENVELOPE PLEASE Interactive Guide to Academy Awards & Oscars*
-.101 http://www.hype.com/movies/oscars/home.htm	*You predict the Oscars for the 68th Annual Academy Awards!*

For the community labeled "b", the top five hub pages include the YAHOO and Zia pages for

Movies/Awards; the top five authorities are as follows:

(+movies +awards) Authorities: 37^{th} non-principal vector, positive end
.223 http://www.bafta.org/ *Croeso i Bafta Cymru - Welcome to Bafta Cymru*
.211 http://www.choiceawards.com/
.211 http://www.sunflower.org/~henryj/movie.htm
.211 http://www.razzies.com *The Golden Raspberry Award Foundation (The "Razzies")*
.211 http://www.emerson.edu/acadepts/mc/EVVY_HP.HTML

Let us return to the issue of ordering the community scores. Since our purpose is only to cover some of the most basic possibilities, our discussion here will be brief. Recall that each community C has an m-coordinate score $a_C = (\alpha_{s_1}(C), \ldots, \alpha_{s_m}(C))$. One natural method would be to sum all the coordinate values in a_C. However, this seems not to be a robust approach, for the reason that certain terms appear to exhibit much greater variance than others in the number of times they occur. Such terms could then wield too great an influence in the scoring function. (Considering the plot in Figure 5, the right-most point would be ranked highest in this measure on the basis of its x-coordinate alone.)

One simple proposal that we feel is borne out better, both intuitively and in our experiments, is the following. For a vector a, define $\mu(a)$ to be the minimum of its coordinate values, and rank the communities according the values of $\mu(a_C)$. In this way, one tries explicitly to find communities in which *all* query terms occur as much as possible. For the above example, the μ-optimal community is the one labeled "b." In Figure 6, we show the results of ranking based on μ for a subset of the topics from the YAHOO *Movies/Genres* list. The 39 communities associated with the first 20 eigenvectors were considered; the second column gives the number of maximal communities, and the third column gives the number lying on the positive hull. Overall, the lexical scoring approach discovered relevant communities associated with weaker eigenvectors in several of the cases.

query	# maxima	pos. hull	μ-optimal community
(+movies +awards) (20 communities)	6	2	Includes Excite's `www.socal.com` *Awards and Festivals* page.
(+movies +classic)	5	2	Some silent and classic movie authorities.
(+movies +comedy)	1	1	General movie resources.
(+movies +documentaries)	5	2	Includes YAHOO's *Documentaries* page.
(+movies +horror)	3	3	6 of 10 are on specific topic; rest are general movie pages.
(+movies +western)	4	3	Includes Zia's *Westerns* page.

Figure 6: Optimal communities for *Entertainment/Movies/Genres* under μ function.

Ultimately, we feel that the problem of scoring and ranking the communities produced by our algorithm, based on association with a query term, presents a number of interesting issues for further work.

7 Conclusion

At various points in the paper, we have discussed some of the basic issues at work in our algorithm, and some of the interesting directions for future work. In this final section, we summarize what we consider to be some of the most substantial of these directions.

(i) We observed at the outset — and the description of the method should make this clear — that the approach we describe need not be restricted to hypermedia. At the most basic level, one can investigate its application to the cross-referencing structure of collections of scientific papers or patents; we have begun experimenting with this method on a large corpus of U.S. patents [19]. But more generally, there are a number of naturally arising directed graph structures in which one can find clear interpretations for the notions of "hubs" and "authorities," and the dense communities which they comprise. Consider, for example, the implicit analogy drawn in [29] between the relationships among modules in a large software system and the basic measures used in bibliometrics. One can also consider the use of our method on graphs defined by financial transactions, or communications (e.g. electronic mail), among a large set of individuals and organizations.

(ii) In the query-based experiments of Section 2, we used a very basic technique for constructing the "one-step neighborhood" T. It is natural to consider more sophisticated methods of defining a neighborhood for the root set provided by the search engine. One promising notion is that of adaptively updating the set as the algorithm proceeds. One could, for example, consider running multiple successive phases of the algorithm — at the end of each phase, one attempts to identify a good set of nodes to use as a root set in the next phase. Here, the notion of "a good set of nodes" is clearly what contains all the complexity. Presumably this judgment would involve both the hub and authority weights, and — to maintain relevance to the initial query — some re-introduction of the textual content of the pages. This, indeed, brings us to our third point.

(iii) One of the striking features of the query-based version of the algorithm is the frequency with which it remains focused on the initial query topic, when this topic is sufficiently "broad." More work, both analytical and experimental, needs to be done in order to better understand the boundary between the set of queries on which the algorithm remains focused, and the set on which it will *diffuse*. For queries on which the pure algorithm does not remain focused, one often obtains a set of generalizations of the initial topic; we are interested in determining methods of focusing the search at various levels of specificity in the resulting communities. Presumably this must make some use of the textual content of the documents that make up these communities. Our experiments in Section 6 indicate some of the basic approaches that are possible in this direction; but there is clearly a range of further techniques that could be investigated.

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. 6th International World Wide Web Conference*, 1997.

[2] A.E. Bayer, J.C. Smart, G.W. McLaughlin, "Mapping intellectual structure of scientific subfields through author co-citations," *J. American Soc. Info. Sci.*, 41(1990), pp. 444–452.

[3] R. Botafogo, E. Rivlin, B. Shneiderman, "Structural analysis of hypertext: Identifying hierarchies and useful metrics," *ACM Trans. Inf. Sys.*, 10(1992), pp. 142–180.

[4] J. Carrière, R. Kazman, "WebQuery: Searching and visualizing the Web through connectivity," *Proc. 6th International World Wide Web Conference*, 1997.

[5] C. Chekuri, M. Goldwasser, P. Raghavan and E. Upfal "Web search using automated classification," submitted for publication.

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

[7] 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), pp. 938–944.

[8] B. Duffy, J. Yacovissi, "Seven self-contradicting reasons why the World Wide Web is such a big deal," *Multimedia Monitor*, August 1996. Also at `http://www.strcom.com/7reasons.htm`.

[9] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, "Indexing by latent semantic analysis," *J. American Soc. Info. Sci.*, 41(1990), pp. 391–407.

[10] Excite Inc. *Excite navigation service*, `http://www.excite.com`.

[11] M. Fielder, "Algebraic connectivity of graphs," *Czech. Math. J.*, 23(1973), pp. 298–305.

[12] FindLaw, *FindLaw – LawCrawler*, `http://www.lawcrawler.com`.

[13] M.E. Frisse, "Searching for information in a hypertext medical handbook," *Communications of the ACM*, 31(7), pp. 880–886.

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

[15] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, 178(1972), pp. 471–479.

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

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

[18] Infoseek Corporation, *Infoseek search engine*, `http://www.infoseek.com`.

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

[20] M.M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, 14(1963), pp. 10–25.

[21] T.R. Kochtanek, "Document clustering using macro retrieval techniques," *J. American Soc. Info. Sci.*, 34(1983), pp. 356–359.

[22] R. Larson, "Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace," *Ann. Meeting of the American Soc. Info. Sci.*, 1996.

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

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

[25] C.J. van Rijsbergen, *Information Retrieval*, Butterworths, 1979. Also at `http://dcs.glasgow.ac.uk/Keith/Preface.html`.

[26] E. Rivlin, R. Botafogo, B. Shneiderman, "Navigating in hyperspace: designing a structure-based toolbox," *Communications of the ACM*, 37(2), 1994, pp. 87–96.

[27] R. Rousseau, G. Van Hooydonk, "Journal production and journal impact factors," *J. American Soc. Info. Sci.*, 47(1996), pp. 775–780.

[28] G. Salton. *Automatic Text Processing.* Addison-Wesley, Reading, MA, 1989.

[29] R.W. Schwanke, M.A. Platoff, "Cross references are features," in *Machine Learning: From Theory to Applications*, S.J. Hanson, W. Remmele, R.L. Rivest, eds., Springer, 1993.

[30] 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), pp. 669–675.

[31] 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), pp. 676–684.

[32] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. American Soc. Info. Sci.*, 24(1973), pp. 265–269.

[33] E. Spertus, "ParaSite: Mining structural information on the Web," *Proc. 6th International World Wide Web Conference*, 1997.

[34] D. Spielman, S. Teng, "Spectral partitioning works: Planar graphs and finite-element meshes," *Processedings of the 37th IEEE Symposium on Foundations of Computer Science*, 1996.

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

[36] R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, D.K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proceedings of the Seventh ACM Conference on Hypertext*, 1996.

[37] H.D. White, K.W. McCain, "Bibliometrics," in *Ann. Rev. Info. Sci. and Technology*, Elsevier, 1989, pp. 119-186.

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

[39] *Zia*, `http://www.zia.com`.