Authoritative Sources in a Hyperlinked Environment*

Jon M. Kleinberg †

Abstract

The network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. We develop a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of contexts on the World Wide Web. The central issue we address within our framework is the distillation of broad search topics, through the discovery of "authoritative" information sources on such topics. We propose and test an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of "hub pages" that join them together in the link structure. Our formulation has connections to the eigenvectors of certain matrices associated with the link graph; these connections in turn motivate additional heuristics for link-based analysis.

*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.

†Dept. 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 network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. In this work, we develop a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of contexts on the World Wide Web (WWW) [4]. In particular, we focus on the use of links for analyzing the collection of pages relevant to a broad search topic, and for discovering the most "authoritative" pages on such topics.

While our techniques are not specific to the WWW, we find the problems of search and structural analysis particularly compelling in the context of this domain. The WWW is a hypertext corpus of enormous complexity, and it continues to expand at a phenomenal rate. Moreover, it can be viewed as an intricate form of populist hypermedia, in which millions of on-line participants, with diverse and often conflicting goals, are continuously creating hyperlinked content. Thus, while individuals can impose order at an extremely local level, its global organization is utterly unplanned — high-level structure can emerge only through *a posteriori* analysis.

Our work originates in the problem of *searching* on the WWW, which we could define roughly as the process of discovering pages that are relevant to a given query. The *quality* of a search method necessarily requires human evaluation, due to the subjectivity inherent in notions such as *relevance*. 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 sizable portion of the WWW and respond on the order of seconds. Although there would be considerable utility in a search tool with a longer response time, 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 objective functions that are both concretely defined *and* correspond to human notions of quality.

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 is more than one type of query, and the handling of each may require different techniques. Consider, for example, the following types of queries.

- *Specific queries.* E.g., "Does Netscape support the JDK 1.1 code-signing API?"
- *Broad-topic queries.* E.g., "Find information about the Java programming language."
- *Similar-page queries.* E.g., "Find pages 'similar' to `java.sun.com`."

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 often difficult to determine the identity of these pages.

For *broad-topic queries*, on the other hand, one expects to find many thousand relevant pages on the WWW; such a set of pages might be generated by variants of term-matching (e.g. one enters a string such as "Gates," "search engines," or "censorship" into a search engine such as AltaVista [17]), 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.* To provide effective search methods under these conditions, one needs a way to filter, from among a huge collection of relevant pages, a small set of the most "authoritative" or "definitive" ones.

This notion of *authority*, relative to a broad-topic query, serves as a central focus in our work. One of the fundamental obstacles we face in addressing this issue is that of accurately modeling authority in the context of a particular query topic. Given a particular page, how do we tell whether it is authoritative?

It is useful to discuss some of the complications that arise here. First, consider the natural goal of reporting `www.harvard.edu`, the home page of Harvard University, as one of the most authoritative pages for the query `"Harvard"`. Unfortunately, there are over a million pages on the WWW that use the term "Harvard," and `www.harvard.edu` is not the one that uses the term most often, or most prominently, or in any other way that would favor it under a text-based ranking function. Indeed, one suspects that there is no purely *endogenous* measure of the page that would allow one to properly assess its authority. Second, consider the problem of finding the home pages of the main WWW search engines. One could begin from the query `"search engines"`, but there is an immediate difficulty in the fact that many of the natural authorities (*Yahoo!*, Excite, AltaVista) do not use the term on their pages. This is a fundamental and recurring phenomenon — as another example, there is no reason to expect the home pages of Honda or Toyota to contain the term "automobile manufacturers."

Analysis of the Link Structure. Analyzing the hyperlink structure among WWW pages gives us a way to address many of the difficulties discussed above. Hyperlinks encode a considerable amount of latent human judgment, and we claim that this type of judgment is precisely what is needed to formulate a notion of authority. Specifically, the creation of a link on 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. Moreover, links afford us the opportunity to find potential authorities purely through the pages that point to them; this offers a way to circumvent the problem, discussed above, that many prominent pages are not sufficiently self-descriptive.

Of course, there are a number of potential pitfalls in the application of links for such a purpose. First of all, links are created for a wide variety of reasons, many of which have nothing to do with the conferral of authority. For example, a large number of links are created primarily for navigational purposes ("Click here to return to the main menu"); others represent paid advertisements.

Another issue is the difficulty in finding an appropriate balance between the criteria of *relevance* and *popularity*, each of which contributes to our intuitive notion of authority. It is instructive to consider the serious problems inherent in the following simple heuristic for locating authoritative pages: Of all pages containing the query string, return those with the greatest number of in-links. We have already argued that for a great many queries (`"search engines"`, `"automobile manufacturers"`, ...), a number of the most authoritative pages do not contain the associated query string. Conversely, this heuristic would consider a universally popular page such as `www.yahoo.com` or `www.netscape.com` to be highly authoritative with respect to any query string that it contained.

In this work, we propose a link-based model for the conferral of authority, and show how it leads to a method that consistently identifies relevant, authoritative WWW pages for broad search topics. Our model is based on the relationship that exists between the authorities for a topic and those pages that link to many related authorities — we refer to pages of this latter type as *hubs*. We observe that a certain natural type of equilibrium exists between hubs and authorities in the graph defined by the link structure, and we exploit this to develop an algorithm that identifies both types of pages simultaneously. The algorithm operates on *focused subgraphs* of the WWW that we construct from the output of a text-based WWW search engine; our technique for constructing such subgraphs is designed to produce small collections of pages likely to contain the most authoritative pages for a given topic.

Overview. Our approach to discovering authoritative WWW sources is meant to have a *global* nature: We wish to identify the most central pages for broad search topics in the context of the WWW as a whole. Global approaches involve basic problems of representing and filtering large volumes of information, since the entire set of pages relevant to a broad-topic query can have a size in the millions. This is in contrast to *local* approaches that seek to understand the interconnections among the set of WWW pages belonging to a single logical site or intranet; in such cases the amount of data is much smaller, and often a different set of considerations dominates.

It is also important to note the sense in which our main concerns are fundamentally different from problems of *clustering.* Clustering addresses the issue of dissecting a heterogeneous population into sub-populations that are in some way more cohesive; in the context of the WWW, this may involve distinguishing pages related to different meanings or senses of a query term. Thus, clustering is intrinsically different from the issue of distilling broad topics via the discovery of authorities, although a subsequent section will indicate some connections. For even if we were able perfectly to dissect the multiple senses of an ambiguous query term (e.g. "Windows" or "Gates"), we would still be left with the same underlying problem of representing and filtering the vast number of pages that are relevant to *each* of the main senses of the query term.

The paper is organized as follows. Section 2 discusses the method by which we construct a focused subgraph of the WWW with respect to a broad search topic, producing a set of relevant pages rich in candidate authorities. Sections 3 and 4 discuss our main algorithm for identifying hubs and authorities in such a subgraph, and some of the applications of this algorithm. Section 5 discusses the connections with related work in the areas of WWW search, bibliometrics, and the study of social networks. Section 6 describes how an extension of our basic algorithm produces multiple collections of hubs and authorities within a common link structure. Finally, Section 7 investigates the question of how "broad" a topic must be in order for our techniques to be effective, and Section 8 surveys some work that has been done on the evaluation of the method presented here.

2 Constructing a Focused Subgraph of the WWW

We can view any collection V of hyperlinked pages as a directed graph $G = (V, E)$: the nodes correspond to the pages, and a directed edge $(p, q) \in E$ indicates the presence of a link from p to q. We say that the *out-degree* of a node p is the number of nodes it has links to, and the *in-degree* of p is is the number of nodes that have links to it. From a graph G, we can isolate small regions, or *subgraphs*, in the following way. If $W \subseteq V$ is a subset of the pages, we use $G[W]$ to denote the graph *induced* on W: its nodes are the pages in W, and its edges correspond to all the links between pages in W.

Suppose we are given a broad-topic query, specified by a query string σ. We wish to determine authoritative pages by an analysis of the link structure; but first we must determine the subgraph of the WWW on which our algorithm will operate. Our goal here is to focus the computational effort on relevant pages. Thus, for example, we could restrict the analysis to the set Q_σ of all pages containing the query string; but this has two significant drawbacks. First, this set may contain well over a million pages, and hence entail a considerable computational cost; and second, we have already noted that some or most of the best authorities may not belong to this set.

Ideally, we would like to focus on a collection S_σ of pages with the following properties.

(i) S_σ is relatively small.

(ii) S_σ is rich in relevant pages.

(iii) S_σ contains most (or many) of the strongest authorities.

By keeping S_σ small, we are able to afford the computational cost of applying non-trivial algorithms; by ensuring it is rich in relevant pages we make it easier to find good authorities, as these are likely to be heavily referenced within S_σ.

How can we find such a collection of pages? For a parameter t (typically set to about 200), we first collect the t highest-ranked pages for the query σ from a text-based search

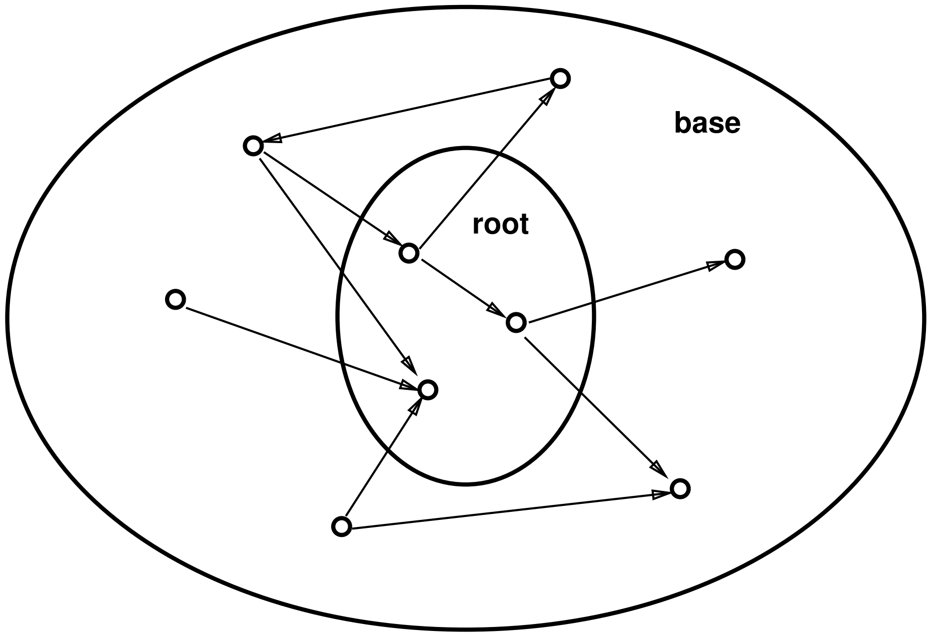

Figure 1: Expanding the root set into a base set.

engine such as AltaVista [17] or Hotbot [57]. We will refer to these t pages as the *root set* R_σ. This root set satisfies (i) and (ii) of the desiderata listed above, but it generally is far from satisfying (iii). To see this, note that the top t pages returned by the text-based search engines we use will all contain the query string σ, and hence R_σ is clearly a subset of the collection Q_σ of *all* pages containing σ. Above we argued that even Q_σ will often not satisfy condition (iii). It is also interesting to observe that there are often extremely few links between pages in R_σ, rendering it essentially "structureless". For example, in our experiments, the root set for the query `"java"` contained 15 links between pages in different domains; the root set for the query `"censorship"` contained 28 links between pages in different domains. These numbers are typical for a variety of the queries tried; they should be compared with the $200 \cdot 199 = 39800$ potential links that could exist between pages in the root set.

We can use the root set R_σ, however, to produce a set of pages S_σ that will satisfy the conditions we're seeking. Consider a strong authority for the query topic — although it may well not be in the set R_σ, it is quite likely to be *pointed to* by at least one page in R_σ. Hence, we can increase the number of strong authorities in our subgraph by expanding R_σ along the links that enter and leave it. In concrete terms, we define the following procedure.

Subgraph(σ,$\mathcal{E}$,t,d)
σ: a query string.
$\mathcal{E}$: a text-based search engine.
t, d: natural numbers.
Let R_σ denote the top t results of $\mathcal{E}$ on σ.

Set $S_\sigma := R_\sigma$
For each page $p \in R_\sigma$
  Let $\Gamma^+(p)$ denote the set of all pages p points to.
  Let $\Gamma^-(p)$ denote the set of all pages pointing to p.
  Add all pages in $\Gamma^+(p)$ to S_σ.
  If $|\Gamma^-(p)| \le d$ then
    Add all pages in $\Gamma^-(p)$ to S_σ.
  Else
    Add an arbitrary set of d pages from $\Gamma^-(p)$ to S_σ.
End
Return S_σ

Thus, we obtain S_σ by growing R_σ to include any page pointed to by a page in R_σ and any page that points to a page in R_σ — with the restriction that we allow a single page in R_σ to bring at most d pages pointing to it into S_σ. This latter point is crucial since a number of WWW pages are pointed to by several hundred thousand pages, and we can't include all of them in S_σ if we wish to keep it reasonably small.

We refer to S_σ as the *base set* for σ; in our experiments we construct it by invoking the `Subgraph` procedure with the search engine AltaVista, $t = 200$, and $d = 50$. We find that S_σ typically satisfies points (i), (ii), and (iii) above — its size is generally in the range 1000-5000; and, as we discussed above, a strong authority need only be referenced by any one of the 200 pages in the root set R_σ in order to be added to S_σ.

In the next section, we describe our algorithm to compute hubs and authorities in the base set S_σ. Before turning to this, we discuss a heuristic that is very useful for offsetting the effect of links that serve purely a navigational function. First, let $G[S_\sigma]$ denote, as above, the subgraph induced on the pages in S_σ. We distinguish between two types of links in $G[S_\sigma]$. 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" here, we mean here the first level in the URL string associated with a page. Since intrinsic links very often exist purely to allow for navigation of the infrastructure of a site, they 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 $G[S_\sigma]$, keeping only the edges corresponding to transverse links; this results in a graph G_σ.

This is a very simple heuristic, but we find it effective for avoiding many of the pathologies caused by treating navigational links in the same way as other links. There are other simple heuristics that can be valuable for eliminating links that do not seem intuitively to confer authority. One that is worth mentioning is based on the following observation. Suppose a large number of pages from a single domain all point to a single page p. Quite often this corresponds to a mass endorsement, advertisement, or some other type of "collusion" among the referring pages — e.g. the phrase "This site designed by ..." and a corresponding link at the bottom of each page in a given domain. To eliminate this phenomenon, we can fix a parameter m (typically $m \approx$ 4-8) and only allow up to m pages from a single domain to point to any given page p. Again, this can be an effective heuristic in some cases, although we did not employ it when running the experiments that follow.

3 Computing Hubs and Authorities

The method of the previous section provides a small subgraph G_σ that is relatively focused on the query topic — it has many relevant pages, and strong authorities. We now turn to the problem of extracting these authorities from the overall collection of pages, purely through an analysis of the link structure of G_σ.

The simplest approach, arguably, would be to order pages by their *in-degree* — the number of links that point to them — in G_σ. We rejected this idea earlier, when it was applied to the collection of *all* pages containing the query term σ; but now we have explicitly constructed a small collection of relevant pages containing most of the authorities we want to find. Thus, these authorities both belong to G_σ and are heavily referenced by pages *within* G_σ.

Indeed, the approach of ranking purely by in-degree does typically work much better in the context of G_σ than in the earlier settings we considered; in some cases, it can produce uniformly high-quality results. However, the approach still retains some significant problems. For example, on the query `"java"`, the pages with the largest in-degree consisted of `www.gamelan.com` and `java.sun.com`, together with pages advertising for Caribbean vacations, and the home page of Amazon Books. This mixture is representative of the type of problem that arises with this simple ranking scheme: While the first two of these pages should certainly be viewed as "good" answers, the others are not relevant to the query topic; they have large in-degree but lack any thematic unity. The basic difficulty this exposes is the inherent tension that exists within the subgraph G_σ between strong authorities and pages that are simply "universally popular"; we expect the latter type of pages to have large in-degree regardless of the underlying query topic.

One could wonder whether circumventing these problems requires making further use of the textual content of pages in the base set, rather than just the link structure of G_σ. We now show that this is not the case — it is in fact possible to extract information more effectively from the links — and we begin from 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. (A skeletal example is depicted in Figure 2; in reality, of course, the picture is not nearly this clean.)

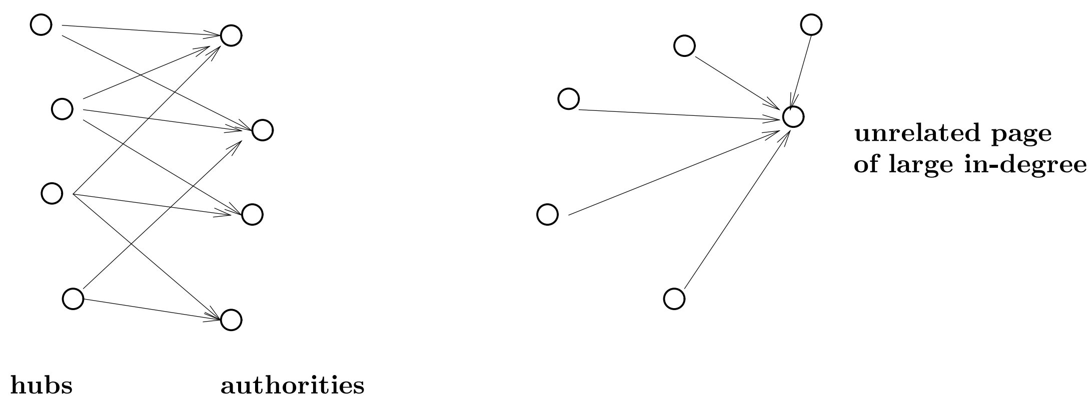

Figure 2: A densely linked set of hubs and authorities.

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 subgraph G_σ, we need a method for breaking this circularity.

An Iterative Algorithm. We make use of the relationship between hubs and authorities via an iterative algorithm that maintains and updates numerical weights for each page. Thus, with each page p, we associate a non-negative *authority weight* $x^{\langle p \rangle}$ and a non-negative *hub weight* $y^{\langle p \rangle}$. We maintain the invariant that the weights of each type are normalized so their squares sum to 1: $\sum_{p \in S_\sigma} (x^{\langle p \rangle})^2 = 1$, and $\sum_{p \in S_\sigma} (y^{\langle p \rangle})^2 = 1$. We view the pages with larger x- and y-values as being "better" authorities and hubs respectively.

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

$$x^{\langle p \rangle} \leftarrow \sum_{q:(q,p) \in E} y^{\langle q \rangle}.$$

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

$$y^{\langle p \rangle} \leftarrow \sum_{q:(p,q) \in E} x^{\langle q \rangle}.$$

Thus $\mathcal{I}$ and $\mathcal{O}$ are the basic means by which hubs and authorities reinforce one another. (See Figure 3.)

Now, to find the desired "equilibrium" values for the weights, 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. We represent the set of weights $\{x^{\langle p \rangle}\}$ as a

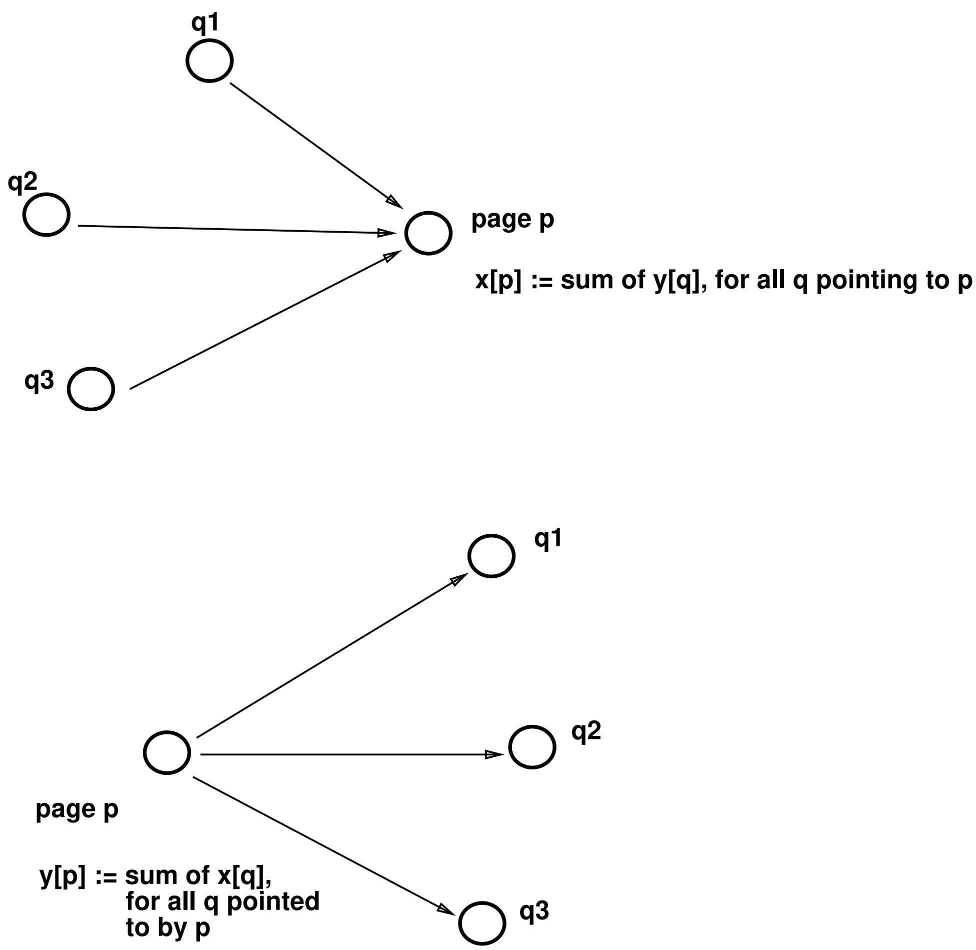

Figure 3: The basic operations.

vector x with a coordinate for each page in G_σ; analogously, we represent the set of weights $\{y^{\langle p \rangle}\}$ as a vector y.

Iterate(G,k)
  G: a collection of n linked pages
  k: a natural number
  Let z denote the vector $(1, 1, 1, \ldots, 1) \in \mathbf{R}^n$.
  Set $x_0 := z$.
  Set $y_0 := z$.
  For $i = 1, 2, \ldots, k$
    Apply the $\mathcal{I}$ operation to (x_{i-1}, y_{i-1}), obtaining new x-weights x'_i.
    Apply the $\mathcal{O}$ operation to (x'_i, y_{i-1}), obtaining new y-weights y'_i.
    Normalize x'_i, obtaining x_i.
    Normalize y'_i, obtaining y_i.
  End
  Return (x_k, y_k).

This procedure can be applied to filter out the top c authorities and top c hubs in the following simple way.

> Filter(G,k,c)
> G: a collection of n linked pages
> k,c: natural numbers
> $(x_k, y_k) :=$ `Iterate`(G, k).
> Report the pages with the c largest coordinates in x_k as authorities.
> Report the pages with the c largest coordinates in y_k as hubs.

We will apply the `Filter` procedure with G set equal to G_σ, and typically with $c \approx$ 5-10. To address the issue of how best to choose k, the number of iterations, we first show that as one applies `Iterate` with arbitrarily large values of k, the sequences of vectors $\{x_k\}$ and $\{y_k\}$ converge to fixed points x^* and y^*.

We require the following notions from linear algebra, and refer the reader to a text such as [30] for more comprehensive background. Let M be a symmetric $n \times n$ matrix. An *eigenvalue* of M is a number λ with the property that, for some vector ω, we have $M\omega = \lambda\omega$. The set of all such ω is a subspace of $\mathbf{R}^n$, which we refer to as the *eigenspace* associated with λ; the dimension of this space will be referred to as the *multiplicity* of λ. It is a standard fact that M has at most n distinct eigenvalues, each of them a real number, and the sum of their multiplicities is exactly n. We will denote these eigenvalues by $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$, indexed in order of decreasing absolute value, and with each eigenvalue listed a number of times equal to its multiplicity. For each distinct eigenvalue, we choose an orthonormal basis of its eigenspace; considering the vectors in all these bases, we obtain a set of eigenvectors $\omega_1(M), \omega_2(M), \ldots, \omega_n(M)$ that we can index in such a way that $\omega_i(M)$ belongs to the eigenspace of $\lambda_i(M)$.

For the sake of simplicity, we will make the following technical assumption about all the matrices we deal with:

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

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

We now prove that the `Iterate` procedure converges as k increases arbitrarily.

Theorem 3.1 *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. Let $G = (V, E)$, with $V = \{p_1, p_2, \ldots, p_n\}$, 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_k is the unit vector in the direction of $(A^T A)^{k-1} A^T z$, and y_k is the unit vector in the direction of $(AA^T)^k z$.

Now, a standard result of linear algebra (e.g. [30]) 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 $M^k v$ converges to $\omega_1(M)$ as k 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 $\{y_k\}$ converges to a limit y^*. Similarly, one can show that if $\lambda_1(A^T A) \neq 0$ (as dictated by Assumption (†)), then $A^T z$ is not orthogonal to $\omega_1(A^T A)$. It follows that the sequence $\{x_k\}$ converges to a limit x^*. ■

The proof of Theorem 3.1 yields the following additional result (in the above notation).

Theorem 3.2 *(Subject to Assumption (†).)* x^* *is the principal eigenvector of* $A^T A$*, and* y^* *is the principal eigenvector of* AA^T*.*

In our experiments, we find that the convergence of `Iterate` is quite rapid; one essentially always finds that $k = 20$ is sufficient for the c largest coordinates in each vector to become stable, for values of c in the range that we use. Of course, Theorem 3.2 shows that one can use any eigenvector algorithm to compute the fixed point x^* and y^*; we have stuck to the above exposition in terms of the `Iterate` procedure 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 compute weights $\{x^{\langle p \rangle}\}$ and $\{y^{\langle p \rangle}\}$ by starting from any initial vectors x_0 and y_0, and performing a fixed bounded number of $\mathcal{I}$ and $\mathcal{O}$ operations.

Basic Results. We now give some sample results obtained via the algorithm, using some of the queries discussed in the introduction.

(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) Authorities

.378	http://www.eff.org/	*EFFweb - The Electronic Frontier Foundation*
.344	http://www.eff.org/blueribbon.html	*The Blue Ribbon Campaign for Online Free Speech*
.238	http://www.cdt.org/	*The Center for Democracy and Technology*
.235	http://www.vtw.org/	*Voters Telecommunications Watch*
.218	http://www.aclu.org/	*ACLU: American Civil Liberties Union*

("search engines") Authorities

.346	http://www.yahoo.com/	*Yahoo!*

.291	http://www.excite.com/	*Excite*
.239	http://www.mckinley.com/	*Welcome to Magellan!*
.231	http://www.lycos.com/	*Lycos Home Page*
.231	http://www.altavista.digital.com/	*AltaVista: Main Page*

(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 R_σ was `www.roadahead.com/`, under the query `"Gates"`; it was ranked 123$^{\text{rd}}$ by AltaVista. This is natural in view of the fact that many of these pages do not contain any occurrences of the initial query string.

It is worth reflecting on two additional points here. First, our only use of the textual content of pages was in the initial "black-box" call to a text-based search engine, which produced the root set R_σ. Following this, the analysis ignored the textual content of pages. The point we wish to make here is not that text is best ignored in searching for authoritative pages; there is clearly much that can be accomplished through the integration of textual and link-based analysis, and we will be commenting on this in a subsequent section. However, the results above show that a considerable amount can be accomplished through essentially a "pure" analysis of link structure.

Second, for many broad search topics, our algorithm produces pages that can legitimately be considered authoritative with respect to the WWW as a whole, despite the fact that it operates without direct access to large-scale index of the WWW. Rather, its only "global" access to the WWW is through a text-based search engine such as AltaVista, from which it is very difficult to directly obtain reasonable candidates for authoritative pages on most queries. What the results imply is that it is possible to reliably estimate certain types of global information about the WWW using only a standard search engine interface; a global analysis of the full WWW link structure can be replaced by a much more local method of analysis on a small focused subgraph.

4 Similar-Page Queries

The algorithm developed in the preceding section can be applied to another type of problem — that of using link structure to infer a notion of "similarity" among pages. Suppose we have found a page p that is of interest — perhaps it is an authoritative page on a topic of interest — and we wish to ask the following type of question: What do users of the WWW consider to be related to p, when they create pages and hyperlinks?

If p is highly referenced page, we have a version of the Abundance Problem: the surrounding link structure will implicitly represent an enormous number of independent opinions about the relation of p to other pages. Using our notion of hubs and authorities, we can provide an approach to the issue of page similarity, asking: In the local region of the link structure near p, what are the strongest authorities? Such authorities can potentially serve as a broad-topic summary of the pages related to p.

In fact, the method of Sections 2 and 3 can be adapted to this situation with essentially no modification. Previously, we initiated our search with a query string σ; our request from the underlying search engine was "Find t pages containing the string σ." We now begin with a page p and pose the following request to the search engine: "Find t pages pointing to p." Thus, we assemble a *root set* R_p consisting of t pages that point to p; we grow this into a base set S_p as before; and the result is a subgraph G_p in which we can search for hubs and authorities.

Superficially, the set of issues in working with a subgraph G_p are somewhat different from those involved in working with a subgraph defined by a query string; however, we find that most of the basic conclusions we drew in the previous two sections continue to apply. First, we observe that ranking pages of G_p by their in-degrees is still not satisfactory; consider for example the results of this heuristic when the initial page p was `www.honda.com`, the home page of Honda Motor Company.

http://www.honda.com	*Honda*
http://www.ford.com/	*Ford Motor Company*
http://www.eff.org/blueribbon.html	*The Blue Ribbon Campaign for Online Free Speech*
http://www.mckinley.com/	*Welcome to Magellan!*
http://www.netscape.com	*Welcome to Netscape*
http://www.linkexchange.com/	*LinkExchange — Welcome*
http://www.toyota.com/	*Welcome to @Toyota*
http://www.pointcom.com/	*PointCom*
http://home.netscape.com/	*Welcome to Netscape*
http://www.yahoo.com	*Yahoo!*

In many cases, the top hubs and authorities computed by our algorithm on a graph of the form G_p can be quite compelling. We show the top authorities obtained when the initial page p was `www.honda.com` and `www.nyse.com`, the home page of the New York Stock Exchange.

(www.honda.com) Authorities

.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

.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`.

5 Connections with Related Work

The analysis of link structures with the goal of understanding their social or informational organization has been an issue in a number of overlapping areas. In this section, we review some of the approaches that have been proposed, divided into three main areas of focus. First, and most closely related to our work here, we discuss research on the use of a link structure for defining notions of *standing*, *impact*, and *influence* — measures with the same motivation as our notion of authority. We then discuss other ways in which links have been integrated into hypertext and WWW search techniques. Finally, we review some work that has made use of link structures for explicit clustering of data.

Standing, Impact, and Influence

Social Networks. The study of social networks has developed several ways to measure the relative *standing* — roughly, "importance" — of individuals in an implicitly defined network. We can represent the network, as above, by a graph $G = (V, E)$; an edge (i, j) corresponds roughly to an "endorsement" of j by i. This is in keeping with the intuition we have already invoked regarding the role of WWW hyperlinks as conferrors of authority. Links may have different (non-negative) *weights*, corresponding to the strength of different endorsements; let A denote the matrix whose $(i, j)^{\text{th}}$ entry represents the strength of the endorsement from a node $i \in V$ to a node $j \in V$.

Katz [35] proposed a measure of standing based on path-counting, a generalization of ranking based on in-degree. For nodes i and j, let $P_{ij}^{\langle r \rangle}$ denote the number of paths of length exactly r from i to j. Let $b < 1$ be a constant chosen to be small enough that $Q_{ij} = \sum_{r=1}^{\infty} b^r P_{ij}^{\langle r \rangle}$ converges for each pair (i, j). Now Katz defines s_j, the *standing* of node j, to be $\sum_i Q_{ij}$ — in this model, standing is based on the total number of paths terminating at node j, weighted by an exponentially decreasing damping factor. It is not difficult to obtain a direct matrix formulation of this measure: s_j is proportional to the j^{th} column sum of the matrix $(I - bA)^{-1} - I$, where I denotes the identity matrix and all entries of A are 0 or 1.

Hubbell [32] proposed a similar model of standing by studying the equilibrium of a certain weight-propagation scheme on nodes of the network. Recall that A_{ij}, the $(i, j)^{\text{th}}$ entry of our matrix A, represents the strength of the endorsement from i to j. Let e_j denote an *a priori* estimate of the standing of node j. Then Hubbell defines the standings $\{s_j\}$ to be a set of values so that the process of endorsement maintains a type of equilibrium — the total "quantity" of endorsement entering a node j, weighted by the standings of the endorsers, is equal to the standing of j. Thus, the standings are the solutions to the system of equations $s_j = e_j + \sum_i A_{ij} s_i$, for $j = 1, \ldots, n$. If e denotes the vector of values $\{e_j\}$, then the vector of standings in this model can be shown to be $(I - A^T)^{-1} e$.

Before discussing the relation of these measures to our work, we consider the way in which they were extended by research in the field of bibliometrics.

Scientific Citations. *Bibliometrics* [22] is the study of written documents and their citation structure. Research in bibliometrics has long been concerned with the use of citations to produce quantitative estimates of the importance and "impact" of individual scientific papers and journals, analogues of our notion of authority. In this sense, they are concerned with evaluating *standing* in a particular type of social network — that of papers or journals linked by citations.

The most well-known measure in this field is Garfield's *impact factor* [26], used to provide a numerical assessment of journals in Journal Citation Reports of the Institute for Scientific Information. Under the standard definition, the impact factor of a journal j in a given year is the average number of citations received by papers published in the previous two years of journal j [22]. Disregarding for now the question of whether two years is the appropriate period of measurement (see e.g. Egghe [21]), we observe that the impact factor is a ranking measure based fundamentally on a pure counting of the in-degrees of nodes in the network.

Pinski and Narin [45] proposed a more subtle citation-based measure of standing, stemming from the observation that not all citations are equally important. They argued that a journal is "influential" if, recursively, it is heavily cited by other influential journals. One can recognize a natural parallel between this and our self-referential construction of hubs and authorities; we will discuss the connections below. The concrete construction of Pinski and Narin, as modified by Geller [27], is the following. The measure of standing of journal

j will be called its *influence weight* and denoted w_j. The matrix A of connection strengths will have entries specified as follows: A_{ij} denotes the fraction of the citations from journal i that go to journal j. Following the informal definition above, the influence of j should be equal to the sum of the influences of all journals citing j, with the sum weighted by the amount that each cites j. Thus, the set of influence weights $\{w_j\}$ is designed to be a non-zero, non-negative solution to the system of equations $w_j = \sum_i A_{ij} w_i$; and hence, if w is the vector of influence weights, one has $w \geq 0$, $w \neq 0$, and $A^T w = w$. This implies that w is a principal eigenvector of A^T. Geller [27] observed that the influence weights correspond to the stationary distribution of the following random process: beginning with an arbitrary journal j, one chooses a random reference that has appeared in j and moves to the journal specified in the reference. Doreian [19, 20] showed that one can obtain a measure of standing that corresponds very closely to influence weights by repeatedly iterating the computation underlying Hubbell's measure of standing: In the first iteration one computes Hubbell standings $\{s_j\}$ from the *a priori* weights $\{e_j\}$; the $\{s_j\}$ then become the *a priori* estimates for the next iteration. Finally, there was been work aimed at the troublesome issue of how to handle journal self-citations (the diagonal elements of the matrix A); see e.g. de Solla Price [15] and Noma [42].

Let us consider the connections between this previous work and our algorithm to compute hubs and authorities. We also begin by observing that pure in-degree counting, as manifested by the impact factor, is too crude a measure for our purposes, and we seek a type of link-based equilibrium among relative node rankings. But the World Wide Web and the scientific literature are governed by very different principles, and this contrast is nicely captured in the distinction between Pinski-Narin influence weights and the hub/authority weights that we compute. Journals in the scientific literature have, to a first approximation, a common purpose, and traditions such as the peer review process typically ensure that highly authoritative journals on a common topic reference one another extensively. Thus it makes sense to consider a one-level model in which authorities directly endorse other authorities. The WWW, on the other hand, is much more heterogeneous, with WWW pages serving many different functions — individual AOL subscribers have home pages, and multinational corporations have home pages. Moreover, for a wide range of topics, the strongest authorities consciously do not link to one another — consider, for example, the home pages of search engines and automobile manufacturers listed above. Thus, they can only be connected by an intermediate layer of relatively anonymous hub pages, which link in a correlated way to a thematically related set of authorities; and our model for the conferral of authority on the WWW takes this into account. This two-level pattern of linkage exposes structure among both the set of hubs, who may not know of one another's existence, and the set of authorities, who may not wish to acknowledge one another's existence.

Hypertext and WWW Rankings. There have been several approaches to ranking pages in the context of hypertext and the WWW. In work predating the emergence of the WWW, Botafogo, Rivlin, and Shneiderman [7] worked with focused, stand-alone hypertext environments. They defined the notions of *index nodes* and *reference nodes* — an index node is one whose out-degree is significantly larger than the average out-degree, and a reference node is one whose in-degree is significantly larger than the average in-degree. They also proposed measures of *centrality* based on node-to-node distances in the graph defined by the link structure.

Carrière and Kazman [9] proposed a ranking measure on WWW pages, for the goal of re-ordering search results. The rank of a page in their model is equal to the sum of its in-degree and its out-degree; thus, it makes use of a "directionless" version of the WWW link structure.

Both of these approaches are based principally on counting node degrees, parallel to the structure of Garfield's *impact factor*. In contrast, Brin and Page [8] have recently proposed a ranking measure based on a node-to-node weight-propagation scheme and its analysis via eigenvectors. Specifically, they begin from a model of a user randomly following hyperlinks: at each page, the user either selects an outgoing link uniformly at random, or (with some probability $p < 1$) jumps to a new page selected uniformly at random from the entire WWW. The stationary probability of node i in this random process will correspond to the "rank" of i, referred to as its *page-rank*.

Alternately, one can view page-ranks as arising from the equilibrium of a process analogous to that used in the definition of the Pinski-Narin influence weights, with the incorporation of a term that captures the "random jump" to a uniformly selected page. Specifically, assuming the WWW contains n pages, letting A denote the $n \times n$ *adjacency matrix* of the WWW, and letting d_i denote the *out-degree* of node i, the probability of a transition from page i to page j in the Brin-Page model is seen to be equal to $A'_{ij} = pn^{-1} + (1-p)d_i^{-1}A_{ij}$. Let A' denote the matrix whose entries are A'_{ij}. The vector of ranks is then a non-zero, non-negative solution to $(A')^T r = r$, and hence it corresponds to the principal eigenvector of $(A')^T$.

One of the main contrasts between our approach and the page-rank methodology is that — like Pinski and Narin's formulation of influence weights — the latter is based on a model in which authority is passed directly from authorities to other authorities, without interposing a notion of hub pages. Brin and Page's use of random jumps to uniformly selected pages is a way of dealing with the resulting problem that many authorities are essentially "dead-ends" in their conferral process.

It is also worth noting a basic contrast in the application of these approaches to WWW search. In [8], the page-rank algorithm is applied to compute ranks for all the nodes in a 24 million page index of the WWW; these ranks are then used to order the results of *subsequent* text-based searches. Our use of hubs and authorities, on the other hand, proceeds without direct access to a WWW index; in response to a query, our algorithm *first* invokes a text-based search and then computes numerical scores for the pages in a relatively small subgraph constructed from the initial search results.

Other Link-Based Approaches to WWW Search

Frisse [25] considered the problem of document retrieval in singly-authored, stand-alone works of hypertext. He proposed basic heuristics by which hyperlinks can enhance notions of *relevance* and hence the performance of retrieval heuristics. Specifically, in his framework, the relevance of a page in hypertext to a particular query is based in part on the relevance of the pages it links to. Marchiori's HyperSearch algorithm [39] is based on such a methodology applied to WWW pages: A relevance score for a page p is computed by a method that incorporates the relevance of pages reachable from p, diminished by a damping factor that decays exponentially with distance from p.

In our construction of focused subgraphs from search engine results in Section 2, the underlying motivation ran also in the opposite direction. In addition to looking at where a page p *pointed* to increase our understanding of its contents, we implicitly used the text on pages that *pointed to* p. (For if pages in the root set for "search engines" pointed to `www.yahoo.com`, then we included `www.yahoo.com` in our subgraph.) This notion is related to that of searching based on *anchor text*, in which one treats the text surrounding a hyperlink as a descriptor of the page being pointed to when assessing the relevance of that page. The use of anchor text appeared in one of the oldest WWW search engines, McBryan's World Wide Web Worm [40]; it is also used in [8, 11, 10].

Another direction of work on the integration of links into WWW search is the construction of search formalisms capable of handling queries that involve predicates over both text and links. Arocena, Mendelzon, and Mihaila [1] have developed a framework supporting WWW queries that combines standard keywords with conditions on the surrounding link structure.

Clustering of Link Structures

Link-based clustering in the context of bibliometrics, hypertext, and the WWW has focused largely on the problem of decomposing an *explicitly represented* collection of nodes into "cohesive" subsets. As such, it has mainly been applied to moderately size sets of objects — for example, a focused collection of scientific journals, or the set of pages on a single WWW site. Earlier we indicated a sense in which the issues we study here are fundamentally different from those encountered in this type of clustering: Our primary concern is that of representing an enormous collection of pages *implicitly*, through the construction of hubs and authorities for this collection. We now discuss some of the prior work on citation-based and hypertext clustering so as to better elucidate its connections to the techniques we develop here. In particular, this will also be useful in Section 6 when we discuss methods for computing multiple sets of hubs and authorities within a single link structure; this can be viewed as a way of representing multiple, potentially very large clusters implicitly.

At a very high level, clustering requires an underlying *similarity function* among objects, and a method for producing *clusters* from this similarity function. Two basic similarity functions on documents to emerge from the study of bibliometrics are *bibliographic coupling* (due to Kessler [36]) and *co-citation* (due to Small [52]). For a pair of 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. Co-citation has been used as a measure of the similarity of WWW pages by Larson [37] and by Pitkow and Pirolli [47]. Weiss et al. [56] define linked-based similarity measures for pages in a hypertext environment that generalize *co-citation* and *bibliographic coupling* to allow for arbitrarily long chains of links.

Several methods have been proposed in this context to produce clusters from a set of nodes annotated with such similarity information. Small and Griffith [54] use breadth-first search to compute the connected components of the undirected graph in which two nodes are joined by an edge if and only if they have a positive co-citation value. Pitkow and Pirolli [47] apply this algorithm to study the link-based relationships among a collection of WWW pages.

One can also use principal components analysis [31, 34] and related dimension-reduction techniques such as multidimensional scaling to cluster a collection of nodes. In this framework, one begins with a matrix M containing the similarity information between pairs of nodes, and a representation (based on this matrix) of each node i as a high-dimensional vector $\{v_i\}$. One then uses the first few non-principal eigenvectors of the similarity matrix M to define a low-dimensional subspace into which the vectors $\{v_i\}$ can be projected; a variety of geometric or visualization-based techniques can be employed to identify dense clusters in this low-dimensional space. Standard theorems of linear algebra (e.g. [30]) in fact provide a precise sense in which projection onto the first k eigenvectors produces the *minimum* distortion over all k-dimensional projections of the data. Small [53], McCain [41], and others have applied this technique to journal and author co-citation data. The application of dimension-reduction techniques to cluster WWW pages based on co-citation has been employed by Larson [37] and by Pitkow and Pirolli [47].

The clustering of documents or hyperlinked pages can of course rely on combinations of textual and link-based information. Combinations of such measures have been studied by Shaw [50, 51] in the context of bibliometrics. More recently, Pirolli, Pitkow, and Rao [46] have used a combination of link topology and textual similarity to group together and categorize pages on the WWW.

Finally, we discuss two other general eigenvector-based approaches to clustering that have been applied to link structures. The area of *spectral graph partitioning* was initiated by the work of Donath and Hoffman [18] and Fiedler [23]; see the recent book by Chung [12] for an overview. Spectral graph partitioning methods relate sparsely connected partitions of an *undirected* graph G to the eigenvalues and eigenvectors of its adjacency matrix A. Each eigenvector of A has a single coordinate for each node of G, and thus can be viewed as an assignment of weights to the nodes of G. Each non-principal eigenvector has both positive and negative coordinates; one fundamental heuristic to emerge from the study of these spectral methods is that the nodes corresponding to the large positive coordinates of a given eigenvector tend to be very sparsely connected to the nodes corresponding to the large negative coordinates of the same eigenvector.

In a different direction, *centroid scaling* is a clustering method designed for representing two types of objects in a common space [38]. Consider, for example, a set of people who have provided answers to the questions of a survey — one may wish to represent both the people and the possible answers in a common space, in a way so that each person is "close" to the answers he or she chose; and each answer is "close" to the people that chose it. Centroid scaling provides an eigenvector-based method for accomplishing this. In its formulation, it thus resembles our definitions of hubs and authorities, which used an eigenvector approach to produce related sets of weights for two distinct types of objects. A fundamental difference, however, is that centroid scaling methods are typically not concerned with interpreting only the largest coordinates in the representations they produce; rather, the goal is to infer a notion of similarity among a set of objects by geometric means. Centroid scaling has been applied to citation data by Noma [43], for jointly clustering citing and cited documents. In the context of information retrieval, the *Latent Semantic Indexing* methodology of Deerwester et al. [16] applied a centroid scaling approach to a vector-space model of documents [48, 49]; this allowed them to represent terms and documents in a common low-dimensional space, in which natural geometrically defined clusters often separate multiple senses of a query term.

6 Multiple Sets of Hubs and Authorities

The algorithm in Section 3 is, in a sense, finding the most *densely* linked collection of hubs and authorities in the subgraph G_σ. defined by a query string σ. There are a number of settings, however, in which one may be interested in finding several densely linked collections of hubs and authorities among the same set S_σ of pages. Each such collection could potentially be relevant to the query topic, but they could be well-separated from one another in the graph G_σ for a variety of reasons. For example,

(1) The query string σ may have several very different meanings. E.g. `"jaguar"` (a useful example we learned from Chandra Chekuri [13]).

(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".

In each of these examples, the relevant documents can be naturally grouped into several clusters. The issue in the setting of broad-topic queries, however, is not simply how to achieve a dissection into reasonable clusters; one must also deal with this in the presence of the Abundance Problem. Each cluster, in the context of the full WWW, is enormous, and so we require a way to distill a small set of hubs and authorities out of each one. We can thus view such collections of hubs and authorities as implicitly providing broad-topic summaries of a collection of large clusters that we never explicitly represent. At a very high level, our motivation in this sense is analogous to that of an information retrieval technique such as *Scatter/Gather* [14], which seeks to represent very large document clusters through text-based methods.

In section 3, we related the hubs and authorities we computed to the principal eigenvectors of the matrices A^TA and AA^T, where A is the adjacency matrix of G_σ. The non-principal eigenvectors of A^TA and AA^T provide us with a natural way to extract additional densely linked collections of hubs and authorities from the base set S_σ. We begin by noting the following basic fact.

Proposition 6.1 *AA^T and A^TA have the same multiset of eigenvalues, and their eigenvectors can be chosen so that $\omega_i(AA^T) = A\omega_i(A^TA)$.*

Thus, each pair of eigenvectors $x_i^* = \omega_i(A^TA)$, $y_i^* = \omega_i(AA^T)$, related as in Proposition 6.1, 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 $\mathcal{I} \cdot \mathcal{O}$ (resp. $\mathcal{O} \cdot \mathcal{I}$) 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, unlike the principal eigenvector, the non-principal eigenvectors have both positive and negative entries. Hence each pair (x_i^*, y_i^*) provides us with two densely connected sets of hubs and authorities: those pages that correspond to the c coordinates with the most positive values, and those pages that correspond to the c coordinates with the most negative values. These sets of hubs and authorities have the same intuitive meaning as those produced in Section 3, although the algorithm to find them — based on non-principal eigenvectors — is less clean conceptually than the method of iterated $\mathcal{I}$ and $\mathcal{O}$ operations. Note also that since the extent to which the weights in x_i^* and y_i^* reinforce each other, the hubs and authorities associated with eigenvectors of larger absolute value will typically be "denser" as subgraphs in the link structure, and hence will often have more intuitive meaning.

In Section 5, we observed that spectral heuristics for partitioning undirected graphs [12, 18, 23] have suggested that nodes assigned large positive coordinates in a non-principal eigenvector are often well-separated from nodes assigned large negative coordinates in the same eigenvector. Adapted to our context, which deals with directed rather than undirected graphs, one can ask whether there is a natural "separation" between the two collections of authoritative sources associated with the same non-principal eigenvector. We will see that in some cases there is a distinction between these two collections, in a sense that has meaning for the query topic. It is worth noting here that the *signs* of the coordinates in any non-principal eigenvector represents a purely arbitrary resolution of the following symmetry: if x_i^* and y_i^* are eigenvectors associated with λ_i, then so are $-x_i^*$ and $-y_i^*$.

Basic Results. We now give some examples of the way in which the application of non-principal eigenvectors produces multiple collections of hubs and authorities. 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 collection of hubs and authorities will often be generated by several of the strongest non-principal eigenvectors. (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 collections of hubs and authorities 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 collections from among the first few non-principal eigenvectors.

We issue the first query as `"jaguar*"`, simply as one way to search for either the word or its plural. For this query, the strongest collections of authoritative sources concerned the Atari Jaguar product, the NFL football team from Jacksonville, and the automobile.

(jaguar*) 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: 2^{nd} 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: 3^{rd} 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 collections of hubs and authorities 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: 1^{st} 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 1^{st} 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 4^{th} 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 also encounter examples where pages from the positive and negative ends of the same non-principal eigenvector exhibit a natural separation. 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 a division between pro-choice and pro-life authorities. The issue is complicated by the existence of hub pages that link extensively to pages from both sides; but in fact the 2^{nd} non-principal eigenvector produces a very clear separation:

(abortion) Authorities: 2^{nd} 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: 2^{nd} 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/	*Pro-Life Advocate*
-.144	http://www.clark.net/pub/jeffd/factbot.html	*The Right Side of the Web*
-.144	http://www.catholic.net/HyperNews/get/abortion.html	

7 Diffusion and Generalization

Let us return to the method of Section 3, in which we identified a single collection of hubs and authorities in the subgraph G_σ associated with a query string σ. The algorithm computes a densely linked collection of pages without regard to their contents; the fact that these pages are relevant to the query topic in a wide range of cases is based on the way in which we construct the subgraph G_σ, ensuring that it is rich in relevant pages. We can view the issue as follows: Many different topics are represented in G_σ, and each is centered around a competing collection of densely linked hubs and authorities. Our method of producing a focused subgraph G_σ aims at ensuring that the most relevant such collection is also the "densest" one, and hence will be found by the method of iterated $\mathcal{I}$ and $\mathcal{O}$ operations.

When the initial query string σ specifies a topic that is not sufficiently broad, however, there will often not be enough relevant pages in G_σ from which to extract a sufficiently dense subgraph of relevant hubs and authorities. As a result, authoritative pages corresponding to competing, "broader" topics will win out over the pages relevant to σ, and be returned by the algorithm. In such cases, we will say that the process has *diffused* from the initial query.

Although it limits the ability of our algorithm to find authoritative pages for narrow or specific query topics, diffusion can be an interesting process in its own right. In particular, the broader topic that supplants the original, too-specific query σ very often represents a natural generalization of σ. As such, it provides a simple way of abstracting a specific query topic to a broader, related one.

Consider, for example, the query `"WWW conferences"`. At the time we tried this query, AltaVista indexed roughly 300 pages containing the string; however, the resulting subgraph G_σ contained pages concerned with a host of more general WWW-related topics, and the main authorities were in fact very general WWW resources.

("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*

In the context of similar-page queries, a query that is "too specific" corresponds roughly to a page p that does not have sufficiently high in-degree. In such cases, the process of diffusion can also provide a broad-topic summary of more prominent pages related to p. Consider, for example, the results when p was `sigact.acm.org`, the home page of the ACM Special Interest Group on Algorithms and Computation Theory, which focuses on theoretical computer science.

(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*

The problem of returning more specific answers in the presence of this phenomenon is the subject of on-going work; in Sections 8 and 9, we briefly discuss current work on the use of textual content for the purpose of focusing our approach to link-based analysis [6, 10, 11]. The use of non-principal eigenvectors, combined with basic term-matching, can be a simple way to extract collections of authoritative pages that are more relevant to a specific query topic. For example, consider the following fact: Among the sets of hubs and authorities corresponding to the first 20 non-principal eigenvectors, the one in which the pages collectively contained the string `"WWW conferences"` the most was the following.

("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/WWWWWW.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*

8 Evaluation

The evaluation of the methods presented here is a challenging task. First, of course, we are attempting to define and compute a measure, "authority," that is inherently based on human judgment. Moreover, the nature of the WWW adds complexity to the problem of evaluation — it is a new domain, with a shortage of standard benchmarks; the diversity of authoring styles is much greater than for comparable collections of printed, published documents; and it is highly dynamic, with new material being created rapidly and no comprehensive index of its full contents.

In the earlier sections of the paper, we have presented a number of examples of the output from our algorithm. This was both to show the reader the type of results that are produced, and because we believe that there is, and probably should be, an inevitable component of *res ipsa loquitur* in the overall evaluation — our feeling is that many of the results are quite striking at an obvious level.

However, there are also more principled ways of evaluating the algorithm. Since the appearance of the conference version of this paper, three distinct user studies performed by two different groups [6, 10, 11] have helped assess the value of our technique in the context of a tool for locating information on the WWW. Each of these studies used a system built primarily on top of the basic algorithm described here, for locating hubs and authorities in a subgraph G_σ via the methods discussed in Sections 2 and 3. However, each of these systems also employed additional heuristics to further enhance relevance judgments. Most significantly, they incorporated text-based measures such as anchor text scores to weight the contribution of individual links differentially. As such, the results of these studies should not be interpreted as providing a direct evaluation of the pure link-based method described here; rather, they assess its performance as the core component of a WWW search tool.

We briefly survey the structure and results of the most recent of these three user studies, involving the CLEVER system of Chakrabarti et al. [10], and refer the reader to that work for more details. The basic task in this study was *automatic resource compilation* — the construction of lists of high-quality WWW pages related to a broad search topic — and the goal was to see how the output of CLEVER compared to that of a manually generated compilation such as the WWW search service *Yahoo!* [58] for a set of 26 topics.

Thus, for each topic, the output of the CLEVER system was a list of ten pages: its five top hubs and five top authorities. *Yahoo!* was used as the main point of comparison, since its manually compiled resource lists can be viewed as representing judgments of "authority" by the human ontologists who compile them. The top ten pages returned by AltaVista were also selected, so as to provide representative pages produced by a fully automatic text-based search engine. All these pages were collected into a single *topic list* for each topic in the study, without an indication of which method produced which page. A collection of 37 users was assembled; the users were required to be familiar with the use of a Web browser, but were not experts in computer science or in the 26 search topics. The users were then asked to rank the pages they visited from the topic lists as "bad," "fair," "good," or "fantastic," in terms of their utility in learning about the topic. This yielded 1369 responses in all, which were then used to assess the relative quality of CLEVER, *Yahoo!*, and AltaVista on each topic. For approximately 31% of the topics, the evaluations of *Yahoo!* and CLEVER were equivalent to within a threshold of statistical significance; for approximately 50% CLEVER was evaluated higher; and for the remaining 19% *Yahoo!* was evaluated higher.

Of course, it is difficult to draw definitive conclusions from these studies. A service such as *Yahoo!* is indeed providing, by its very nature, a type of human judgment as to which pages are "good" for a particular topic. But even the nature of the quality judgment is not well-defined, of course. Moreover, many of the entries in *Yahoo!* are drawn from outside submissions, and hence represent less directly the "authority" judgments of *Yahoo!*'s staff.

Many of the users in these studies reported that they used the lists as starting points from which to explore, but that they visited many pages not on the original topic lists generated by the various techniques. This is, of course, a natural process in the exploration of a broad topic on the WWW, and the goal of resource lists appears to be generally for the purpose of facilitating this process rather than for replacing it.

9 Conclusion

We have discussed a technique for locating high-quality information related to a broad search topic on the WWW, based on a structural analysis of the link topology surrounding "authoritative" pages on the topic. It is useful to highlight four basic components of our approach.

- For broad topics on the WWW, the amount of relevant information is growing extremely rapidly, making it continually more difficult for individual users to filter the available resources. To deal with this problem, one needs notions beyond those of relevance and clustering — one needs a way to distill a broad topic, for which there may be millions of relevant pages, down to a representation of very small size. It is for this purpose that we define a notion of "authoritative" sources, based on the link structure of the WWW.

- We are interested in producing results that are of as a high a quality as possible in the context of what is available on the WWW *globally*. Our underlying domain is not restricted to a focused set of pages, or those residing on a single Web site.

- At the same time, we infer global notions of structure without directly maintaining an index of the WWW or its link structure. We require only a basic interface to any of a number of standard WWW search engines, and use techniques for producing "enriched" samples of WWW pages to determine notions of structure and quality that make sense globally. This helps to deal with problems of scale in handling topics that have an enormous representation on the WWW.

- We began with the goal of discovering *authoritative pages*, but our approach in fact identifies a more complex pattern of social organization on the WWW, in which hub pages link densely to a set of thematically related authorities. This equilibrium between hubs and authorities is a phenomenon that recurs in the context of a wide variety of topics on the WWW. Measures of impact and influence in bibliometrics have typically lacked, and arguably not required, an analogous formulation of the role that hubs play; the WWW is very different from the scientific literature, and our framework seems appropriate as a model of the way in which authority is conferred in an environment such as the Web.

This work has been extended in a number of ways since its initial conference appearance. In Section 8 we mentioned systems for compiling high-quality WWW resource lists that have been built using extensions to the algorithms developed here; see Bharat and Henzinger [6] and Chakrabarti et al. [10, 11]. The implementation of the Bharat-Henzinger system made use of the recently developed Connectivity Server (Bharat et al. [5]), which provides very efficient retrieval for linkage information contained in the AltaVista index.

With Gibson and Raghavan, we have used the algorithms described here to explore the structure of "communities" of hubs and authorities on the WWW [28]. We find that the notion of topic generalization discussed in Section 7 provides one valuable perspective from which to view the overlapping organization of such communities. In a separate direction, also with Gibson and Raghavan, we have investigated extensions of the present work to the analysis of relational data, and considered a natural, non-linear analogue of spectral heuristics in this setting [29].

There a number of interesting further directions suggested by this research, in addition to the currently on-going work mentioned above. We will restrict ourselves here to three such directions.

First, we have used structural information about the graph defined by the links of the WWW, but we have not made use of its patterns of traffic, and the paths that users implicitly traverse in this graph as they visit a sequence of pages. There are a number of interesting and fundamental questions that can be asked about WWW traffic, involving both the modeling of such traffic and the development of algorithms and tools to exploit information gained from traffic patterns (see e.g. [2, 3, 33]). It would be interesting to ask how the approach developed here might be integrated into a study of user traffic patterns on the WWW.

Second, the power of eigenvector-based heuristics is not something that is fully understood at an analytical level, and it would be interesting to pursue this question in the context of the algorithms presented here. One direction would be to consider random graph models that contain enough structure to capture certain global properties of the WWW, and yet are simple enough so that the application of our algorithms to them could be analyzed. More generally, the development of clean yet reasonably accurate random graph models for the WWW could be extremely valuable for the understanding of a range of link-based algorithms. Some work of this type has been undertaken in the context of the *latent semantic indexing* technique in information retrieval [16]: Papadimitriou et al. [44] have provided a theoretical analysis of latent semantic indexing applied to a basic probabilistic model of term use in documents. In another direction, motivated in part by our work here, Frieze, Kannan, and Vempala have analyzed sampling methodologies capable of approximating the singular value decomposition of a large matrix very efficiently [24]; understanding the concrete connections between their work and our sampling methodology in Section 2 would be very interesting.

Finally, the further development of link-based methods to handle information needs other than broad-topic queries on the WWW poses many interesting challenges. As noted above, work has been done on the incorporation of textual content into our framework as a way of "focusing" a broad-topic search [6, 10, 11], but one can ask what other basic informational structures one can identify, beyond hubs and authorities, from the link topology of hypermedia such as the WWW. The means by which interaction with a link structure can facilitate the discovery of information is a general and far-reaching notion, and we feel that it will continue to offer a range of fascinating algorithmic possibilities.

Acknowledgements

In the early stages of this work, I benefited enormously from discussions with Prabhakar Raghavan and with Robert Kleinberg; I thank Soumen Chakrabarti, Byron Dom, David Gibson, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins for on-going collaboration on extensions and evaluations of this work; and I thank Rakesh Agrawal, Tryg Ager, Rob Barrett, Marshall Bern, Tim Berners-Lee, Ashok Chandra, Monika Henzinger, Alan Hoffman, David Karger, Lillian Lee, Nimrod Megiddo, Christos Papadimitriou, Peter Pirolli, Ted Selker, Eli Upfal, and the anonymous referees of this paper, 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] R. Barrett, P. Maglio, D. Kellem, "How to personalize the Web," *Proc. Conf. on Human Factors in Computing Systems*, 1997.

[3] O. Berman, M.J. Hodgson, D. Krass. "Flow-interception problems," in *Facility Location: A Survey of Applications and Methods*, Z. Drezner, Ed., Springer 1995.

[4] T. Berners-Lee, R. Cailliau, A. Luotonen, H.F. Nielsen, A. Secret. "The World-Wide Web," *Communications of the ACM*, 37(1994), pp. 76–82.

[5] K. Bharat, A. Broder, M.R. Henzinger, P. Kumar, S. Venkatasubramanian, "Connectivity Server: Fast Access to Linkage Information on the Web," *Proc. 7th Intl. World Wide Web Conf.*, 1998.

[6] K. Bharat, M.R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," *Proc. ACM Conf. Res. and Development in Information Retrieval*, 1998.

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

[8] S. Brin, L. Page, "Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. 7th International World Wide Web Conference*, 1998.

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

[10] S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Experiments in Topic Distillation," *ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, 1998.

[11] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," *Proc. 7th International World Wide Web Conference*, 1998.

[12] F.R.K. Chung, *Spectral Graph Theory*, AMS Press, 1997.

[13] C. Chekuri, M. Goldwasser, P. Raghavan and E. Upfal "Web search using automated classification," poster at *6th International World Wide Web Conference*, 1997.

[14] D.R. Cutting, J. Pedersen, D.R. Karger, J.W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections," *Proc. ACM Conf. Res. and Development in Information Retrieval*, 1992.

[15] D. de Solla Price, "The analysis of square matrices of scientometric transactions," *Scientometrics*, 3(1981), pp. 55–63.

[16] 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.

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

[18] W.E. Donath, A.J. Hoffman, "Lower bounds for the partitioning of graphs", *IBM Journal of Research and Development*, 17(1973).

[19] P. Doreian, "Measuring the relative standing of disciplinary journals," *Inf. Proc. and Management*, 24(1988), pp. 45–56.

[20] P. Doreian, "A measure of standing for citation networks within a wider environment," *Inf. Proc. and Management*, 30(1994), pp. 21–31.

[21] L. Egghe, "Mathematical relations between impact factors and average number of citations," *Inf. Proc. and Management*, 24(1988), pp. 567–576.

[22] L. Egghe, R. Rousseau, *Introduction to Informetrics*, Elsevier, 1990.

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

[24] A. Frieze, R. Kannan, S. Vempala, "Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations," *Proc. 39th IEEE Symp. on Foundations of Computer Science*, 1998.

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

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

[27] N. Geller, "On the citation influence methodology of Pinski and Narin," *Inf. Proc. and Management*, 14(1978), pp. 93–95.

[28] D. Gibson, J. Kleinberg, P. Raghavan, "Inferring Web Communities from Link Topology," *Proc. 9th ACM Conference on Hypertext and Hypermedia*, 1998.

[29] D. Gibson, J. Kleinberg, P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," *Proc. 24th Intl. Conf. on Very Large Databases*, 1998.

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

[31] H. Hotelling, "Analysis of a complex statistical variable into principal components," *J. Educational Psychology*, 24(1933), pp. 417–441.

[32] C.H. Hubbell, "An input-output approach to clique identification," *Sociometry*, 28(1965), pp. 377-399.

[33] B. Huberman, P. Pirolli, J. Pitkow, R. Lukose, "Strong Regularities in World Wide Web Surfing," *Science*, 280(1998).

[34] I.T. Jolliffe. Principal Component Analysis. Springer-Verlag, 1986.

[35] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, 18(1953), pp. 39–43.

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

[37] 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.

[38] J.H. Levine, "Joint-space analysis of 'pick-any' data: Analysis of choices from an unconstrained set of alternatives," *Psychometrika*, 44(1979), pp. 85–92.

[39] M. Marchiori, "The quest for correct information on the Web: Hyper search engines," *Proc. 6th International World Wide Web Conference*, 1997.

[40] O. McBryan, "GENVL and WWWW: Tools for taming the Web," *Proc. 1st International World Wide Web Conference*, 1994.

[41] K. McCain, "Co-cited author mapping as a valid representation of intellectual structure," *J. American Soc. Info. Sci.*, 37(1986), pp. 111–122.

[42] E. Noma, "An improved method for analyzing square scientometric transaction matrices," *Scientometrics*, 4(1982), pp. 297–316.

[43] E. Noma, "Co-citation analysis and the invisible college," *J. American Soc. Info. Sci.*, 35(1984), pp. 29–33.

[44] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Proc. ACM Symp. on Principles of Database Systems*, 1998.

[45] G. Pinski, F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," *Inf. Proc. and Management*, 12(1976), pp. 297–312.

[46] 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.

[47] J. Pitkow, P. Pirolli, "Life, death, and lawfulness on the electronic frontier," *Proceedings of ACM SIGCHI Conference on Human Factors in Computing*, 1997.

[48] C.J. van Rijsbergen, *Information Retrieval*, Butterworths, 1979.

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

[50] 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.

[51] 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.

[52] 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.

[53] H. Small, "The synthesis of specialty narratives from co-citation clusters," *J. American Soc. Info. Sci.*, 37(1986), pp. 97–110.

[54] H. Small, B.C. Griffith, "The structure of the scientific literatures I. Identifying and graphing specialties," *Science Studies* 4(1974), pp. 17–40.

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

[56] 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.

[57] Wired Digital, Inc., *Hotbot*, `http://www.hotbot.com`.

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