RJ 10076 (91892) May 29, 1997
Computer Science

Research Report

AUTHORITATIVE SOURCES IN A HYPERLINKED ENVIRONMENT

Jon M. Kleinberg

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, California 95120-6099

RJ 10076 (91892) May 29, 1997
Computer Science

AUTHORITATIVE SOURCES IN A HYPERLINKED ENVIRONMENT

Jon M. Kleinberg*

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, California 95120-6099

ABSTRACT:

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

The central problem we consider is that of determining the relative *authority* of pages in such environments. This issue is central to a number of basic hypertext search tasks; for example, if the result of a query-based search consists of a large set of relevant pages, one may wish to select a small subset of the most "definitive" or "authoritative" pages to present to a user. At the same time, it is clearly difficult to formulate a definition of authority precise enough to be used in such contexts. We propose and test an algorithmic formulation of the notion of authority in a hyperlinked environment, based on its link structure, together with a method for extracting information about the authority of pages from this structure.

*IBM Almaden Research Center, San Jose CA 95120, on leave from Department of Computer Science, Cornell University, Ithaca NY 14853.

1 Introduction

The link structure of a hypermedia environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. Versions of this principle have been studied in the hypertext research community [12, 22, 3, 32] and (in a context predating hypermedia) through journal citation analysis in the field of bibliometrics [33]. But for the problem of searching in hyperlinked environments such as the World Wide Web, it is clear from current techniques that the information inherent in the links has yet to be significantly exploited. In this work we develop a new method for automatically extracting certain types of information about a hypermedia environment from its link structure, and we report on experiments that demonstrate its effectiveness for a variety of search problems on the WWW.

Our methods appear to be applicable fairly broadly, to structures that are implicitly, as well as explicitly, linked. For the purposes of this discussion, and to set up the background for our experiments, we focus on the problem of searching for information in a hyperlinked environment, and in particular on the World Wide Web.

Searching could be defined as the process of discovering pages that are relevant to a given query. Of course, this definition contains several loaded terms; to begin with, notions such as *relevance* necessarily defy precise formulation in the absence of human judgment. Additionally, it seems best not to take too unified a view of the notion of a *query*: there are many possible types of queries that one might wish to pose, and it is likely that the proper handling of each type requires a different set of techniques. Consider, for example, the following types of queries.

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

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

For *broad-topic queries*, on the other hand, one could easily expect to find several thousand relevant pages in an environment such as the WWW; such a set of pages might be generated by variants of term-matching (e.g. one enters a string such as "summer olympics," "web browsers," or "affirmative action" into a search engine such as AltaVista [6]), or by more sophisticated means. Thus, there is not an issue of scarcity here. Instead, the fundamental difficulty lies in what could be called the *Abundance Problem*: *The number of pages that could reasonably be returned as "relevant" is far too large for a human user to digest.* Thus, to provide effective methods for automated search under these constraints, one does not necessarily need stronger versions of classical information retrieval notions such as relevance; rather one needs a method of providing a user, from a large set of relevant pages, a small collection of the most "authoritative" or "definitive" ones.

Our work here is centered around these issues raised by the Abundance Problem, and particularly around the problem of discovering the most authoritative pages in a large hyperlinked environment. The problem is particularly interesting in that much of its complexity has nothing to do with the "search" component; rather, we face the dilemma that in order to search for authoritative pages, one must first formulate a concrete means of *recognizing* them. Unfortunately, "authority" is perhaps an even more nebulous concept than "relevance," again highly subject to human judgment; and without a concrete formal model for it, there is no hope of analyzing it algorithmically.

Let us bring the notion of links back into this picture. We claim that an environment such as the WWW is explicitly annotated with precisely the type of human judgment that we need in order to formulate a notion of authority. Specifically, the creation of a link in the WWW represents a concrete indication of the following type of judgment: the creator of page p, by including a link to page q, has in some measure *conferred authority* on q. Thus we are faced with the following natural problem: given the vast size of the underlying environment, can we synthesize the highly unreliable information contained in the presence of individual links in a way that provides an estimate of which are the truly "authoritative" pages?

Our Approach. We propose a concrete formulation of the notion of *authority* in a hyperlinked environment, based on its link structure, together with a method for extracting information about the authority of pages from this structure. Our formulation is built on a mechanism whereby authority is explicitly *conferred* over directed links; we view authority as being transferred iteratively through the link structure, and establish convergence results about the tendency of this process to reach an equilibrium state. Thus, our underlying mechanism serves both as a *definition* of authority, and as the nucleus of an efficient *algorithm* to discover authoritative pages. We will show the application of our algorithm in the context of several search problems on the WWW.

Inferring information from the links of the WWW is, to be sure, fraught with complications. Links are created for a wide variety of reasons, some of which have nothing to do with the conferral of authority. Moreover, it is difficult to rank pages based on incident links in a way that preserves the relevance of the highly-ranked pages to an underlying query topic — one wants to make sure that a single irrelevant page with a huge number of incident links does not steal the authority away from a large cluster of somewhat more sparsely connected relevant pages. Thus we will see that it is critical to develop methods that can perform a *robust* synthesis of the highly nosiy information contained in the existence of a single link, and of the links incident to a single page.

There is one more fundamental component to our approach. One can picture the WWW as a form of "populist hypermedia," with the property that individual users and well-known authorities are simultaneous participants. This leads to the phenomenon that certain pages, created by individuals who are not in themselves "authoritaties," serve as very potent *conferrers* of authority: these are typically large collections of links to resources on topics of interest to the creator of the page. In some sense, they can be viewed as serving a function similar to that of survey papers in the setting of the academic literature, although the quality of the information they provide is much less reliable. We refer to such pages as *hubs*. Hubs are a surprisingly recurrent feature of this type of environment; their distinctive properties

have been observed in a number of sources. For example, Duffy and Yacovissi write

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

Hubs and authorities possess what could be called a *mutually reinforcing relationship*: given a good set of hubs, the pages they point to the most densely are likely to be good authorities; and conversely, given a good set of authorities, the pages that point to the greatest density of them are likely to be good hubs. Our basic method exploits this relationship by an iterative technique that converges simultaneously to sets of hubs and authorities. This is described in Section 2.

Clustering. We have so far been speaking in terms of operations on a set of pages that are (presumed to be) relevant to a unified topic. In many contexts, the set of pages one is dealing with can be naturally *clustered* into several "dense" regions, with the property that the regions are well-separated under some measure. There are many choices for measures of separation under which one can perform clustering; for example, much work in information retrieval has used vector-space methods for clustering based on index term frequencies (see e.g. [24, 21, 9]).

An extension of our basic method provides a technique for using the link structure to discover clusters among a set of pages. There has been previous work on this, both in the context of hypermedia [22, 32] and citation analysis [33, 26]; this will discussed more fully in the section on previous work, below. The clustering provided by our method appears to be quite different from most of these previous approaches. It can be viewed as a form of *spectral graph partitioning*, which has been studied in the context of combinatorial optimization (e.g. [7, 11, 30]), but with the novelty here that it can be used together with the notion of hubs and authorities. Specifically, any pure clustering method suffers from a variant of the basic *Abundance Problem* discussed above: each cluster contains far too many pages to present to a human user. Thus, a method is needed to provide a succinct representation of each cluster that is discovered. Our approach provides a set of hubs and authorities for each cluster, which can then be used as "representative pages" for the entire cluster.

Experiments. In Sections 3 and 4, we report the results of tests of the methodology in the context of the WWW. We believe these experiments indicate several respects in which the definitions formulated here succeed in capturing notions of authority in such a hyperlinked environment.

The main types of experiments that were carried out are the following.

(i) *Broad-topic queries.* A query string is issued to AltaVista, and the first 200 pages returned are used as a *root set* around which to build a subgraph in the WWW. The algorithm is then used to find hubs and authorities in this subgraph.

(ii) *Similar-page queries*. Given a page p, a subgraph consisting of pages within two (undirected) links of p is constructed. The algorithm is then applied to this subgraph.

(iii) *Intranet analysis*. The algorithm is applied to the set of all pages residing on a given server.

For the query-based experiments in (i), note that the algorithm only makes use of the link structure of the resulting subgraph, and hence the query string itself is used solely to generate the intial root set of pages. Of course, we are taking an extreme position in throwing away knowledge of the query string once a root set of pages has been identified. The value of this approach is that it shows the extent to which authoritative pages can be identified even in (essentially) the complete absence of term-based information. It also allows the basic method to avoid many of the typical pitfalls associated with term-based searching.

The primary risk with this approach is clearly that the iterative method for locating hubs and authorities can *diffuse* to pages on a topic different from the original query. In our experiments, this typically happens when the query defines a topic that is relatively specific, and the algorithm converges to pages on a generalization of this topic. We feel that the issue of *diffusion* is a very interesting one, and return to it in Section 5: we investigate some basic techniques for counteracting the effects of *diffusion* via a *post-processing phase* based on term-matching, once a set of clusters has been produced.

1.1 Previous Work

Methodologically, our work is most closely connected to the area of *bibliometrics* [33] — the study of written documents and their citation structure. Some related work has also been done in the hypertext research community. This work has focused predominantly on the use of citations and/or explicit hyperlinks as a means of clustering and enhancing relevance judgments.

Two basic measures of document similarity to emerge from the study of bibliometrics are *bibliographic coupling* [18] and *co-citation* [28]. For two documents p and q, the former quantity is equal to the number of documents cited by both p and q, and the latter quantity is the number of documents that cite both p and q. We will see that these two quantities arise inside the analysis of our method. Shaw [26, 27] uses a combination of these measures, together with some textual measures, as part of a graph-based clustering algorithm. Documents constitute the nodes of an undirected graph whose edge weights are measures of similarity; edges are deleted in order of increasing weight, producing a hierarchical clustering. (See e.g. [21].) Schwanke and Platoff [25] discuss an interesting application of these bibliometric measures for clustering purposes in a completely different realm — that of analyzing the relationships among modules in a large software system.

There has also been work in bibliometrics on using citation counts to assess the "impact" of scientific journals; this is more closely related to the issue we are considering here. The classic work in this area is that of Garfield [14]; see also e.g. [15, 23]. In addition to basic differences in the algorithms themselves, a fundamental difference between our method and the classical bibliometric work lies in our explicit separation of *hubs* and *authorities*, and investigation of the equilibrium that exists between them — as discussed above, this is a phenomenon that perhaps makes more sense in the context of hypermedia such as the WWW,

where one finds several categories of participants, than it does in the classic bibliometric setting of scientific journals, which ostensibly serve a uniform role. At the same time, it has been observed in bibliometric studies that *review journals* often constitute exceptions to certain basic principles [15, 23]; it would be interesting to see whether the distinction between hubs and authorities would be useful in this regard.

There has been some work on using hyperlinks for clustering and searching in the hypertext research community. Rivlin, Botafogo, and Schneiderman [3, 22] use basic graph-theoretic notions such as connectivity, as well as "compactness" measures based on node-to-node distances, to identify clusters in the graph of a hypertext environment. Weiss et al. [32] define similarity measures among pages in a hypertext environment based on the link structure; these measures are generalizations of *co-citation* and *bibliographic coupling* to allow for arbitrarily long chains of references. Larson [20] performs a co-citation analysis of a set of pages relevant to a sample query and generates clusters by dimension-reduction techniques. Finally, Frisse [12] describes a method that is applicable in a tree-structured environment: the *relevance* of a page with respect to a query is also based on the relevance of its descendants in the tree.

More recently, Arocena, Mendelzon, and Mihaila [1] and Spertus [29] have described frameworks for constructing WWW queries from a combination of term-matching and link-based predicates. Finally, in what is perhaps the closest work to ours in the hypertext community, Carrière and Kazman [4] propose a link-based method for visualizing and ranking the results of queries returned by WWW search engines. Their method is essentially to (1) enlarge the set S of returned pages to include any page joined to a member of S via a link (in either direction); and then (2) rank each page p in this enlarged set according to the number of pages connected to p via links (again, in either direction). Although the notion of augmenting search engine results to a "one-step neighborhood" is a basic component of our experiments in Sections 3 and 4, the algorithmic component of our work differs significantly from that of [4]. In particular, we make crucial use of the directionality of hyperlinks, including the explicit distinction of *hubs* and *authorities*; our ranking of pages is not obtained by a direct counting of neighbors in the link structure; and our framework naturally allows for the construction of multiple clusters of ranked pages.

2 The Method

In this section, we give a precise description of our method. The underlying algorithm can be run on an arbitrary hyperlinked set of pages, and it is in this context that we describe it. In the experiments to be described in the following sections, we use several different techniques for constructing hyperlinked environments from subsets of the WWW— it is important to note that the underlying *algorithm* being used in all cases is the same, and the experiments only differ in the means by which the environment is constructed initially.

We represent the environment as a directed graph $G = (V, E)$. The set V consists of the n pages in the environment, and a directed edge $(p, q) \in E$ indicates the presence of a link from p to q. We associate a non-negative *authority weight* x_p and a non-negative *hub weight* y_p with each page $p \in V$. We will maintain the invariant that the weights of each type are

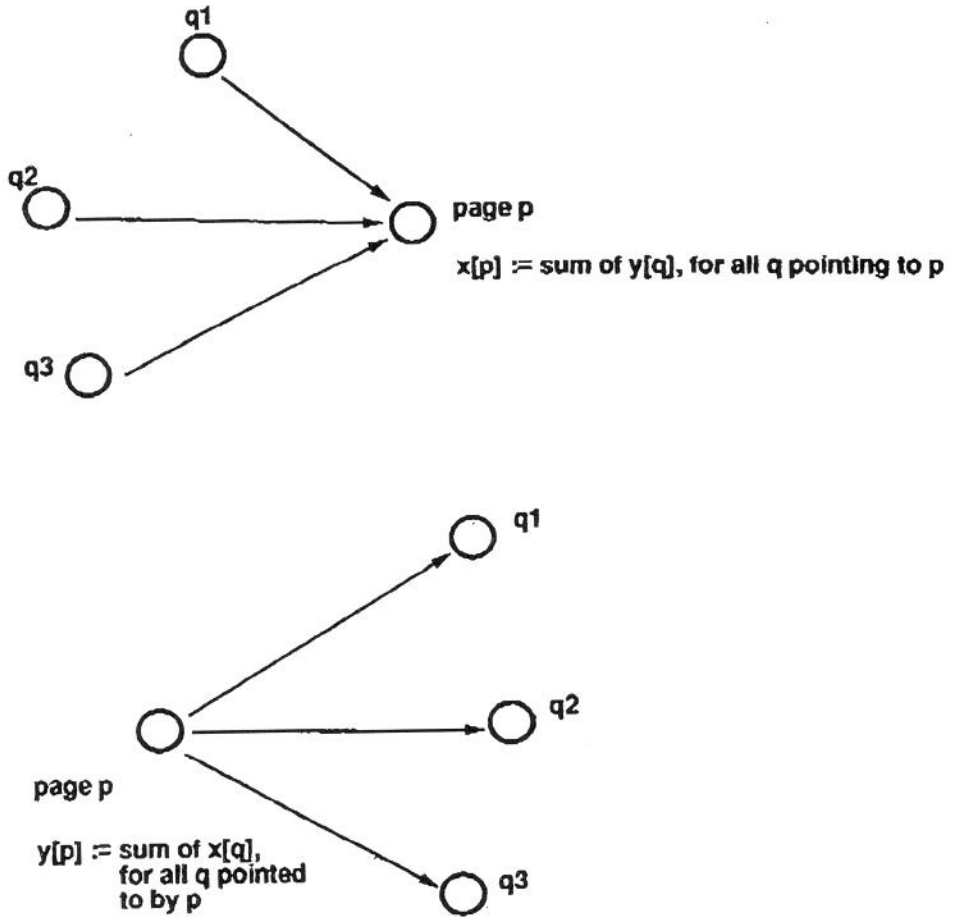

Figure 1: The basic operations.

normalized so that their squares sum to 1:

$$\sum_{p \in V} x_p^2 = 1.$$

$$\sum_{p \in V} y_p^2 = 1.$$

We view the pages with larger x- and y-values as being "better" authorities and hubs respectively.

The method can be developed in terms of two operations on the sets of weights, which we denote by $\mathcal{I}$ and $\mathcal{O}$. An iterative algorithm based on these two operations produces the final hub and authority weights. The result of this iterative algorithm turns out to be equivalent to an eigenvector computation, and from the point of view of computational efficiency it is clearly best treated as such. However, in order to motivate the algorithm itself, we feel it is best explained in terms of iterated $\mathcal{I}$ and $\mathcal{O}$ operations.

2.1 The Basic Operations

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

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

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

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

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

2.2 The Iterative Algorithm

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

Let z denote the vector $(1,1,1,\ldots,1)$.
Initially set $x \leftarrow z$.
$y \leftarrow z$.
For $i = 1,2,3,\ldots$
    Apply the $\mathcal{I}$ operation
    Apply the $\mathcal{O}$ operation
    Normalize x and y
The sequence of (x,y) pairs produced converges to a limit (x^*,y^*)
    (see Theorem 2.1).
Return (x^*,y^*) as the authority and hub weights.

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

$$(x_n, y_n) = (\mathcal{OI})^n(z,z).$$

For the proof, we need the following additional background. For an $n \times n$ symmetric matrix M, let $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$ denote the eigenvalues of M (all of which are real), indexed in order of decreasing absolute value. Let $\omega_i(M)$ denote the eigenvector associated with λ_i. For the sake of simplicity, we will make the following 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 following analysis becomes more elaborate, but it is not affected in any substantial way.

Theorem 2.1 *The sequences $x_1, x_2, x_3, \ldots$ and $y_1, y_2, y_3, \ldots$ converge.*

Proof. Write $V = \{p_1, p_2, \ldots, p_n\}$, and let A denote the *adjacency matrix* of the graph G; that is 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

$$x \leftarrow A^T y$$

$$y \leftarrow Ax$$

respectively. Thus x_n is the unit vector in the direction of $(A^TA)^{n-1}A^Tz$, and y_n is the unit vector in the direction of $(AA^T)^n z$.

Now, a standard result of linear algebra (see e.g. [16]) states that if M is a symmetric $n \times n$ matrix, and v is a vector that is not orthogonal to the principal eigenvector $\omega_1(M)$, then the unit vector in the direction of $M^n v$ converges to $\omega_1(M)$ as n increases without bound. It follows also that if M has only non-negative entries, then the principal eigenvector of M has only non-negative entries.

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

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

Theorem 2.2 *Assume (†) for the matrices A^TA and AA^T. Then x^* is the principal eigenvector of A^TA, and y^* is the principal eigenvector of AA^T.*

Given this, we can write an equivalent version of our basic algorithm.

Form the adjacency matrix A of the graph G.
Return $x^* = \omega_1(A^TA)$ as the authority vector.
Return $y^* = \omega_1(AA^T)$ as the hub vector.

The advantage of this description of the algorithm is that there are generally (iterative) methods to compute eigenvectors that are computationally more efficient than the iterative method implicit in the previous description.

As indicated in Section 1, the output of this process from the user's point of view would be the c pages with the largest x^*-values and the c pages with the largest y^*-values, for a small constant c; this represents the algorithm's estimate of the strongest authorities and hubs among the collection of pages.

Two further remarks are in order. First, one does not have to run the above process to convergence; one can instead compute weight $\{x_p\}$ and $\{y_p\}$ by starting from the "flat"

vector z and performing a fixed bounded number of $\mathcal{I}$ and $\mathcal{O}$ operations. In many of our experiments, even using a small number of iterations gives good results.

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

2.3 Clustering

An extension of our algorithm provides a method for *clustering*. Our definition of *clustering* is, by design, not fully precise; we mean the term to include essentially any representation of a subset of the pages as a collection of sets (clusters), with the property that the density of links *within* each set is non-trivially greater than the density of links *between* different sets.

A well-known heuristic for finding clusters in an undirected graph is *spectral graph partitioning* (see e.g. [7, 11, 30]): one computes a non-principal eigenvector of the adjacency matrix, and clusters the nodes with positive coordinates separately from the nodes with negative coordinates. In many settings, this technique provides very good clustering performance. Let us develop a setting in which this technique can be combined with our notions of hubs and authorities.

First, it is worth noting the following basic fact explicitly.

Theorem 2.3 *AA^T and A^TA have the same multiset of eigenvalues, and their eigenvectors can be chosen so that $\omega_i(AA^T) = A\omega_i(A^TA)$.*

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

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

■

Thus, each pair of eigenvectors $x_i^* = \omega_i(A^TA)$, $y_i^* = \omega_i(AA^T)$, related as in Theorem 2.3, 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, the pair of weights (x_i^*, y_i^*) has precisely the *mutually reinforcing relationship* that we are seeking in authority/hub pairs. Moreover, applying $(\mathcal{IO})$ (resp. $(\mathcal{OI})$) multiplies the magnitude of x_i^* (resp. y_i^*) by a factor of $|\lambda_i|$; thus $|\lambda_i|$ gives precisely the extent to which the hub weights y_i^* and authority weights x_i^* *reinforce* one another.

Our basic clustering method, then, is the following. For each pair (x_i^*, y_i^*) of non-principal eigenvectors, we produce two clusters of hubs and authorities: the most positive entries of x_i^* and y_i^* form one cluster, and the most negative entries form another. There are several concrete ways of implementing this; here is the main one that we adopt in the rest of the paper. Let c be a constant parameter value, and say that a *cluster* is a pair of sets (A, H), each consisting of c pages. (We picture A and H as being a reinforcing pair of authority and hub sets.) From the pair (x_i^*, y_i^*) of non-principal eigenvectors, we produce the following two

clusters: (A_i, H_i), consisting of the c pages with the most positive coordinates in x_i^* and y_i^* respectively; and (A_i', H_i'), consisting of the c pages with the most negative coordinates in x_i^* and y_i^* respectively. Note that the sets A_i, A_i', H_i, H_i' typically do not constitute a partition of the underlying set of pages.

For a cluster (A, H) produced in this way, the extent to which the sets A and H "reinforce" one another is dictated by the absolute value of the eigenvalue λ_i. Thus we say that the two clusters associated with the i^{th} eigenvectors are *stronger* than the two clusters associated with the j^{th} eigenvectors, for $1 < i < j$. (Recall that the eigenvalues are indexed decreasing order of absolute value.) In the experiments, we typically find that most of the clusters are associated with near-zero eigenvalues, and these tend not to provide meaningful information.

For a given i, the heuristic principle underlying spectral partitioning asserts that the authority/hub pair corresponding to the positive entries of (x_i^*, y_i^*) will be "well-separated" in the graph G from the authority/hub pair corresponding to the negative entries of (x_i^*, y_i^*). The extent to which this property appears to hold in practice will be an issue when we discuss experimental results in the following section.

Finally, we note that the pair (x_1^*, y_1^*) of *principal* eigenvectors holds a distinguished position within this framework. Since they have only non-negative coordinates, they cannot be used for clustering in the sense developed here, and so they yield only a single cluster of hubs and authorities, by the basic method of Section 2.2. Since λ_1 is the eigenvalue of maximal absolute value, this *principal cluster* is the "strongest" one, in the terminology above.

3 Experiments: Broad-Topic Queries

Our first experiment proceeds roughly as follows. We supply a query string to a search engine, and obtain a set S of k "relevant" pages, for some parameter k. We then produce an *augmented set* T consisting of S, together with pages that either point to a page in S, or are pointed to by a page in S. Finally, we run our algorithm on the hyperlinked set T.

The reason for working on a set larger than S is simple: S is quite likely not to contain the most authoritative pages on the query topic, since a typical search engine is quite likely not to rank them highly enough. On the other hand, they are very likely to be *referenced* by a reasonable number of the pages in S, and hence appear in the "one-step neighborhood" T.

A more difficult question is the following. Our algorithm makes no use of the terms contained in the pages — so how can we be sure that the authorities we return are relevant to the original query? Of course, we can't be sure (and we discuss this point much more extensively in Section 5); but our experiments have shown that for sufficiently broad queries, it is nearly always the case. The basic argument why we are likely to produce relevant authorities is the following. Since the set T consists only of pages that are linked to S, it is still generally *dense* in pages that are relevant to the query. So although T will also contain heavily referenced pages that are not relevant (e.g. sites such as `www.yahoo.com` turn up very frequently, regardless of the query), the set of authorities is rich in those that are relevant. Hence the pages that point to relevant authorities will acquire hub weight much more rapidly than pages that point to irrelevant authorities, and this concurrently will cause the relevant authorities to gain in authority weight more rapidly.

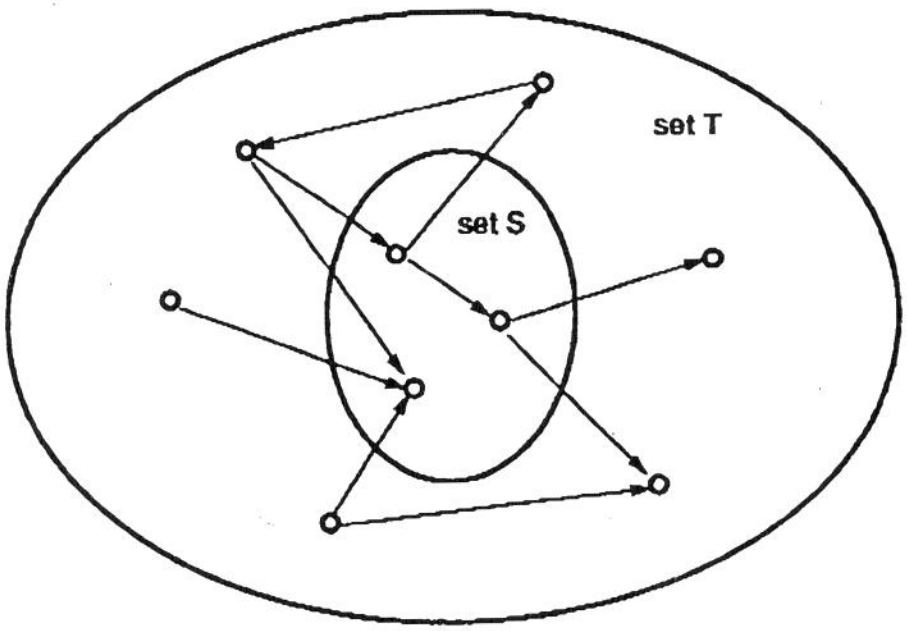

Figure 2: Growing a neighborhood around the root set.

This picture gives a more quantitative sense of what it means for a query to be sufficiently "broad": it should have the property that even after augmenting the set S to its one-step neighborhood T, one still has a density of relevant pages. It also makes clear why it would be less effective, in terms of remaining focused on the query, to arbitrarily enlarge the set T further. (Consider, for example, what would happen if one defined T to be the entire WWW.)

It is worth adding to this discussion that one can clearly consider more sophisticated methods for growing the augmented set T (which, indeed, could be modified adaptively as the algorithm proceeds). This is an interesting direction for future work.

3.1 Description of the Experiment

For the most part, we used the search engine AltaVista [6], with $|S| = 200$. (Below, we discuss some tests in which different search engines were used to provide root sets S for the same query string.) AltaVista also provides the ability to determine the set of pages that have a directed link to a given page p; having this ability is crucial in constructing the set T. In defining the construction of the set T, we wanted to prevent pages of enormous in-degree from "swamping" the resulting graph; thus, we arbitrarily allowed each page in S to bring at most 50 pages that point to it into the set T.

Finally, we distinguished throughout these experiments between two types of links in the WWW. We say that a link is *transverse* if it is between pages with different domain names, and *intrinsic* if it is between pages with the same domain name. By "domain name," we mean here the first level in the URL string associated with a page. Since intrinsic links are very often created simply to help users navigate the infrastructure of a site, they tend to convey much less information than transverse links about the authority of the pages they

point to. Hence, we considered only transverse links in building the graphs and running our algorithm. Finding a more refined handling of intrinsic links, and the information provided by the URL hierarchy more generally, is also an interesting open question.

Thus, in detail, the experiment is the following.

(i) A query string is issued to AltaVista, and a set S of 200 pages is returned.

(ii) A set T is formed consisting of S, any page pointed to by a page in S, and up to 50 pages pointing to each page in S.

(iii) The algorithm of Section 2 is run on the hyperlinked set T, using only transverse links.

3.2 Basic Results

It is a non-trivial problem to find objective means for evaluating our results, and this is essentially for obvious reasons: the definition of "authority" is highly subjective, and there do not currently exist standard benchmarks for the task of finding authoritative sources. Nevertheless, finding ways to determine the effectiveness of our algorithm is a crucial and interesting issue, and we discuss our approach to it here.

Essentially, we focused on the following two types of tests.

(i) Clearly, one method is the following: choose various broad topics for which one knows of obvious authoritative pages, and see whether the algorithm discovers them. Despite the obvious elements of bias in such a test, we feel that it can be quite instructive.

(ii) At a less subjective level, one can make use of searchable hierarchies such as YAHOO [34] and the WWW Virtual Library [31]. These provide both lists of query topics, and human-created lists of authoritative pages on these topics. Thus, they can be used to determine the effectiveness of the algorithm on the query topics that they cover.

Let us first discuss some examples of the first type. Our example on the opening page concerned the query "web browsers" — a query string for which AltaVista locates approximately 50,000 pages. The top-ranked authorities found by our algorithm are the following:

```
("web browsers") Authorities 0
0) 0.225604 http://www.ncsa.uiuc.edu/SDG/Software/WinMosaic/HomePage.html
        NCSA Windows Mosaic Home Page
1) 0.202585 http://home.mcom.com/home/welcome.html
        Welcome to Netscape
2) 0.196804 http://galaxy.einet.net/EINet/EINet.html
        TradeWave Corporation
3) 0.188907 http://www.interport.net/slipknot/slipknot.html
        .... SlipKnot Home Page ....
4) 0.188536 http://galaxy.einet.net/EINet/WinWeb/WinWebHome.html
        winWeb and MacWeb
5) 0.185042 http://www.microsoft.com/ie/
        Microsoft Internet Explorer
```

A word about the format of the results, which we will use throughout. The top line lists the query string provided to the search engine, followed by the index of the eigenvector used to produce the authority weights. The index 0 indicates the cluster associated with the principal eigenvector; the index $+j$ (resp. $-j$) indicates the cluster associated with the positive (resp. negative) entries of the j^{th} non-principal eigenvector. In each successive pair of lines, the first number is the ordinal ranking, and the second number is the authority weight of the associated page. Pages are represented by their URL's, with the title of the page (from the html title field, if present) listed on the following line.

Thus, we note that the above set of authorities includes home pages for NCSA Mosaic, Netscape, and Microsoft Internet Explorer. It is also worth noting that no pages from any of these three domains were included in the initial root set S provided by AltaVista.

The following is a sampling of other basic examples in which the algorithm returned intuitively "accurate" sets of authorities.

```
(java) Authorities 0
0) 0.360910 http://www.gamelan.com/
        Gamelan
1) 0.258038 http://java.sun.com/
        JavaSoft Home Page
2) 0.208786 http://lightyear.ncsa.uiuc.edu/ srp/java/javabooks.html
        The Java Book Pages

(Gates) Authorities 0
0) 0.643337 http://www.roadahead.com/
        Bill Gates:  The Road Ahead
1) 0.458681 http://www.microsoft.com/
        Welcome to Microsoft
2) 0.440370 http://www.microsoft.com/corpinfo/bill-g.htm

(+censorship +net) Authorities 0
0) 0.421033 http://www.eff.org/
        EFFweb - The Electronic Frontier Foundation
1) 0.394431 http://www.eff.org/blueribbon.html
        The Blue Ribbon Campaign for Online Free Speech
2) 0.390115 http://www.cdt.org/
        The Center for Democracy and Technology

("summer olympics") Authorities 0
0) 0.133837 http://www.atlanta.olympic.org/
        Official 1996 Olympic Web Site - Home Page
1) 0.116685 http://www.atlantagames.com/
        Atlanta Games
2) 0.107641 http://www.fanmail.olympic.ibm.com/

(exchanges) Authorities 0
```

```
0) 0.277346 http://www.cme.com/
        Futures and Options at the Chicago Mercantile Exchange
1) 0.208407 http://www.liffe.com/
        Welcome to LIFFE
2) 0.206329 http://www.amex.com/
        The American Stock Exchange - The Smarter Place to Be
```

Among all these pages, the only one which occurred in the initial root set S was `www.roadahead.com/`, under the query (`Gates`); it was ranked 123rd by AltaVista. This should not be so surprising — many of these pages do not contain any occurrences of the initial query term.

Of course, we have also encountered a variety of cases in which the algorithm produced pages having very little to do with original query. Some general phenomena about the behavior of the algorithm can be observed from these cases; we defer this discussion to Section 5.

3.3 Clustering

We also give some basic examples of the use of non-principal eigenvectors for clustering. Even when the set T of pages we construct remains relatively focused on the original query topic, well-separated clusters can emerge within this set for a variety of reasons:

(i) The string has several very different meanings. E.g. (`jaguar jaguars`) [5].

(ii) The string is used by several distinct technical communities. E.g. (`"randomized algorithms"`).

(iii) The string refers to a highly polarized issue, involving groups that are not likely to link to one another. E.g. (`abortion`).

For the first of these queries, the algorithm produced fairly strong clusters centered around the Atari Jaguar system, the Jacksonville Jaguars football team, and the Jaguar automobile. We list them in this order. (Pages concerning the jungle cat were much more weakly represented in the overall set T, and they only emerged in association with much weaker eigenvectors.)

```
(jaguar jaguars) Authorities 0
0) 0.370595 http://www2.ecst.csuchico.edu/~jschlich/Jaguar/jaguar.html
1) 0.347560 http://www-und.ida.liu.se/~t94patsa/jserver.html
2) 0.292649 http://tangram.informatik.uni-kl.de:8001/ rgehm/jaguar.html

(jaguar jaguars) Authorities +2
0) 0.255207 http://www.jaguarsnfl.com/
        Official Jacksonville Jaguars NFL Website
1) 0.137331 http://www.nando.net/SportServer/football/nfl/jax.html
        Jacksonville Jaguars Home Page
2) 0.133919 http://www.ao.net/~brett/jaguar/index.html
```

```
        Brett's Jaguar Page

(jaguar jaguars) Authorities +3
0) 0.227547 http://www.jaguarvehicles.com/
        Jaguar Cars Global Home Page
1) 0.227547 http://www.collection.co.uk/
        The Jaguar Collection - Official Web site
2) 0.211947 http://www.moran.com/sterling/sterling.html
```

For the second of these queries, the strong clusters produced by the algorithm included home pages of theoretical computer scientists, compendia of mathematical software, and pages on wavelets:

```
("randomized algorithms") Authorities +1
0) 0.125326 http://theory.lcs.mit.edu/~goemans/
        Michel X. Goemans
1) 0.122735 http://theory.lcs.mit.edu/~spielman/
        Dan Spielman's Homepage
2) 0.122735 http://www.nada.kth.se/~johanh/
        Johan Hastad

("randomized algorithms") Authorities -1
0) -0.001160 http://lib.stat.cmu.edu/
        StatLib Index
1) -0.001159 http://www.geo.fmi.fi/prog/tela.html
        Tela
2) -0.001077 http://gams.nist.gov/
        GAMS : Guide to Available Mathematical Software

("randomized algorithms") Authorities -4
0) -0.176285 http://www.amara.com/current/wavelet.html
        Amara's  Wavelet Page
1) -0.172956 http://www-ocean.tamu.edu/~baum/wavelets.html
        Wavelet sources
2) -0.161813 http://www.mathsoft.com/wavelets.html
        Wavelet Resources
```

In the third of these queries, the clusters produced by the first non-principal eigenvector were not particularly polarized; this was apparently due to the effect of hub pages that linked extensively to pages on both sides of the issue. However, the second non-principal eigenvector did produce a fairly clear partition between pro-choice and pro-life pages:

```
(abortion) Authorities +2
0) 0.321561 http://www.caral.org/abortion.html
        Abortion and Reproductive Rights Internet Resources
1) 0.219474 http://www.plannedparenthood.org/
```

```
        Welcome to Planned Parenthood
2) 0.195394 http://www.gynpages.com/
        Abortion Clinics OnLine
3) 0.172782 http://www.oneworld.org/ippf/
        IPPF Home Page
4) 0.162707 http://www.prochoice.org/naf/
        The National Abortion Federation
5) 0.161035 http://www.lm.com/~lmann/feminist/abortion.html

(abortion) Authorities -2
0) -0.197192 http://www.awinc.com/partners/bc/commpass/lifenet/lifenet.htm
        LifeWEB
1) -0.169559 http://www.worldvillage.com/wv/square/chapel/xwalk/html/peter.htm
        Healing after Abortion
2) -0.164170 http://www.nebula.net/~maeve/lifelink.html
3) -0.150814 http://members.aol.com/pladvocate/
4) -0.144885 http://www.clark.net/pub/jeffd/factbot.html
        The Right Side of the Web
5) -0.144829 http://www.catholic.net/HyperNews/get/abortion.html
        abortion
```

3.4 Searchable Hierarchies

A more objective method for generating test queries is to use the topic lists from a *searchable hierarchy*. By this we mean a site which indexes a collection of general topics in a tree-structured fashion; the leaves of the tree consist, essentially, of hub pages for individual topics. YAHOO is a basic example of a searchable hierarchy. As alluded to above, basing the experiment on such a searchable hierarchy has two main advantages: it provides an objective means for generating topics, and it provides a pre-made hub page against which to check the set of pages returned.

We tested the algorithm on the topics in *Health/Medicine*, *Science/Physics*, *Entertainment/Movies/Genres*, and several other topic lists in YAHOO. In what follows, we analyze the behavior of the algorithm on ten arbitrarily chosen topics from *Health/Medicine*: acupuncture, anatomy, anesthesiology, audiology, cardiology, dermatology, endocrinology, epidemiology, gastroenterology, and hematology. For the sake of concreteness, we consider only the hubs and authorities determined in the principal cluster.

Although the queries were drawn explicitly from YAHOO, pages from several different searchable hierarchies appeared as high-scoring hub pages for many of the topics. These included both general-purpose hierarchies which attempt to represent all topics, and specialized medical hierarchies which contain pages only for a range of medical key words. The former category contains sites such as YAHOO[34], Galaxy [13], Zia [35], and the distributed WWW Virtual Library [31]; the latter category contains sites such as the MedMark index at `medmark.bit.co.kr`, and the MedWeb index at `www.gen.emory.edu/medweb`.

The table in Figure 3 provides the following information: from among the top 20 hubs and top 20 authorities found for each query term, it lists the set of pages referenced by each

of YAHOO, Galaxy, and Zia. Referenced pages are labeled by their rank, with the prefix 'A' or 'H'; a dash indicates that the hierarchy did not contain a page for the associated topic.

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

Figure 3: Top hubs and authorities referenced by three searchable hierarchies, on ten topics from *Health/Medicine*.

Of the three hierarchies, the relevant page from YAHOO itself was among the top 20 hub pages once (as H9 under (audiology)), the relevant Zia page was among the top 20 hub pages once (as H10 under (audiology)), and the relevant Galaxy page was among the top 20 hub pages twice (as H14 under (gastroenterology) and H10 under (hematology)).

The identity of the top hub page (H0) can be summarized as follows.

- The www Virtual Library hierarchy contained pages for two of the topics — anesthesiology and epidemiology. For both of these topics, it provided the top hub page.
- For five of the queries — audiology, dermatology, endocrinology, gastroenterology, and hematology — the top hub page was the relevant page from the MedMark hierarchy at `medmark.bit.co.kr`.
- For the remaining three topics, the top hub page contained pointers to many pages relevant to the query, but did not seem to belong to a larger searchable hierarchy.

There are a number of things that can be observed from the results.

- First of all, the top hub pages (and authorities) were relevant to the initial query for all ten topics. This can clearly be attributed in some measure to the abundance of professionally assembled hub pages for these topics.
- The general-purposes hierarchies YAHOO, Galaxy, and Zia did not score as highly as some of the focused medical hierarchies, though they consistently referenced authorities

in the top 20. The WWW Virtual Library, while also a general purpose hierarchy, scored extremely well.

- The success of the WWW Virtual Library for the query (epidemiology) had a lot to do with the phenomenon of *mirroring*: many sites maintain local copies of the WWWVL, and each of these has the potential to behave as an independent hub page. In particular, the union of the mirrored pages can wield more influence in the algorithm than a single copy of the page could. In the case of (epidemiology), the first, second, third, fourth, and sixth hubs were all copies of the WWWVL page. The MedMark index also benefitted from mirroring on several of the queries.

  The issue of mirroring is a difficult one to handle. In particular, it is not the case that one can prevent it by looking for textually identical pages with different URL's; in many cases, each mirrored copy has been independently modified at the local site in non-trivial ways, and hence they are far from identical to one another. Again, this appears to be an interesting direction for future work. It is also worth noting that a mirrored page does not always have large influence in the algorithm; for example, multiple copies of the YAHOO pages were found as part of the (anatomy) query, but none of them were among the top 20 hub pages.

3.5 Alternative Root Sets

We have observed in a number of examples that the initial set returned by AltaVista lacked the most authoritative pages, and yet the algorithm still discovered them. This suggests a sort of robustness to the underlying method, in that it can converge to a "good" set of pages despite a highly sub-optimal initial root set. Our intuition is, indeed, that the method should be able to converge to a reasonable answer from essentially a randomly chosen set S of pages containing the initial query term, although we have not attempted to design an experiment that would test this rigorously.

However, we can design an experiment that tests the dependence on the root set S to some extent. In particular, rather than using AltaVista to generate the initial 200 pages, we can use a different search engine such as Infoseek [17] or Excite [10]. This will typically have the effect of changing the root set S considerably. We have run a number of tests of this form; the results appear to indicate that the main clusters remain intact as one varies the root set, although the eigenvectors with which they are associated can change. In particular, the hubs and authorities associated with the principal eigenvector under one root set S can become associated with a non-principal eigenvector under a different root set S'.

We show the top authorities produced for the query ("web browsers") starting from the root sets produced by Infoseek and Excite. The authorities associated with the principal eigenvector in the Infoseek experiment look very similar to those in the AltaVista experiment.

```
("web browsers") (via Infoseek) Authorities 0
0) 0.227512 http://www.ncsa.uiuc.edu/SDG/Software/WinMosaic/HomePage.html
        NCSA Windows Mosaic Home Page
1) 0.197968 http://www.interport.net/slipknot/slipknot.html
        .... SlipKnot Home Page ....
```

```
2) 0.195323 http://galaxy.einet.net/EINet/WinWeb/WinWebHome.html
        winWeb and MacWeb
3) 0.192836 http://www.microsoft.com/
        Welcome to Microsoft
4) 0.185404 http://home.mcom.com/home/welcome.html
        Welcome to Netscape
5) 0.179201 http://www.ncsa.uiuc.edu/General/NCSAHome.html
```

In the Excite experiment, the authorities associated with the principal eigenvector no longer consist of web browser companies; rather, they are centered around a different topic. However, the cluster of browser manufacturers does survive intact, associated with a non-principal eigenvector.

```
("web browsers") (via Excite) Authorities 0
0) 0.111403 http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/fill-out-forms/ov
1) 0.109485 http://hoohoo.ncsa.uiuc.edu/cgi/overview.html
        The Common Gateway Interface
2) 0.106151 http://www.gamelan.com/
        Gamelan
3) 0.103274 http://www.w3.org/hypertext/WWW/MarkUp/html-spec/html-spectoc.html
4) 0.103226 http://worldwidemart.com/scripts/
        Matt's Script Archive
5) 0.103153 http://info.med.yale.edu/caim/StyleManualTop.HTML

("web browsers") (via Excite) Authorities -3
0) -0.560194 http://home.netscape.com/
        Welcome to Netscape
1) -0.252814 http://www.microsoft.com/
        Welcome to Microsoft
2) -0.231935 http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html
        NCSA Windows Mosaic Home Page
3) -0.224464 http://www.microsoft.com/ie/
        Microsoft Internet Explorer
4) -0.215408 http://galaxy.einet.net/EINet/MacWeb/MacWebHome.html
        winWeb and MacWeb
5) -0.189268 http://www.ncsa.uiuc.edu/SDG/Software/MacMosaic/MacMosaicHome.html
        Mosaic for the Macintosh
```

3.6 Cross-Linguistic Queries

Since the algorithm does not look at the text of pages, it can arbitrarily mix pages of different languages within a single unified cluster, provided that the pages densely refer to one another. This suggests that the method has the potential to be effective on *cross-linguistic queries*, in which a query issued in one language produces relevant (or even authoritative) documents in another.

Consider, for example, the first of the topics from YAHOO's *Science/Physics* page:

```
(astrophysics) Authorities 0
0) 0.226805 http://fits.cv.nrao.edu/www/astronomy.html
        AstroWeb:  Astronomy/Astrophyics on the Internet
1) 0.189811 http://cdsweb.u-strasbg.fr/Simbad.html
        The SIMBAD astronomical database
2) 0.189364 http://www.aas.org/
        American Astronomical Society Home Page
3) 0.183309 http://heasarc.gsfc.nasa.gov/
        HEASARC/GSFC Home Page
4) 0.175138 http://adsabs.harvard.edu/abstract-service.html
5) 0.169933 http://cdsweb.u-strasbg.fr/CDS.html
        CDS, Strasbourg
6) 0.161780 http://adswww.harvard.edu/
        The NASA Astrophysics Data System Home Page
```

We now issue the same query in French and German. We discover clusters similar to the above, although now they are associated with non-principal eigenvectors.

```
(astrophysique) Authorities -8
0) -0.253138 http://cdsweb.u-strasbg.fr/CDS.html
        CDS, Strasbourg
1) -0.239516 http://adswww.harvard.edu/
        The NASA Astrophysics Data System Home Page
2) -0.208483 http://cdsweb.u-strasbg.fr/Simbad.html
        The SIMBAD astronomical database
3) -0.189571 http://adsabs.harvard.edu/abstract-service.html
4) -0.153339 http://fits.cv.nrao.edu/www/astronomy.html
        AstroWeb:  Astronomy/Astrophyics on the Internet
5) -0.134335 http://www.hq.eso.org/eso-homepage.html
        European Southern Observatory Homepage
6) -0.122625 http://info.er.usgs.gov/network/science/astronomy/index.html
        Astronomy and Space Science

(astrophysik) Authorities -7
0) -0.306625 http://adswww.harvard.edu/
        The NASA Astrophysics Data System Home Page
1) -0.273664 http://cdsweb.u-strasbg.fr/Simbad.html
        The SIMBAD astronomical database
2) -0.273649 http://adsabs.harvard.edu/abstract-service.html
3) -0.237155 http://aibn55.astro.uni-bonn.de:8000/
4) -0.186439 http://www.univ-rennes1.fr/ASTRO/astro.english.html
        Astronomical pictures & animations - An online 3 gigabytes astronomical
5) -0.173030 http://aorta.tat.physik.uni-tuebingen.de/
6) -0.139660 http://fits.cv.nrao.edu/www/astronomy.html
```

4 Experiments: Similar Pages and Intranet Analysis

We now turn to the other two types of experiments mentioned in Section 1; we consider applications of the basic method in settings where no use is made of query terms.

4.1 Similar-Page Queries

Suppose we have found a page p that is of interest — perhaps it is an authoritative page on a topic of interest — and we want to use the link structure of the environment to discover whether there exist pages that are "similar" to p We show here how a minor modification of the technique from Section 3 provides a link-based definition of page similarity, together with a method for discovering similar pages. It is based on the following notion: If we build an appropriate "neighborhood" T of pages around p, and p turns out to be a good authority for this set T, or for some cluster of it, then the other authorities in the same cluster as p will exhibit a type of linked-based similarity to p. A similar notion holds if p turns out to be a good hub for some cluster of T.

We performed experiments in which the construction of the set T was strictly analogous to the construction in the previous section:

(i) A set S is constructed, consisting of all the pages pointed to by p, and up to 200 pages that point to p.

(ii) A set T is formed consisting of S, any page pointed to by a page in S, and up to 50 pages pointing to each page in S.

(iii) The algorithm of Section 2 is run on the hyperlinked set T, using only transverse links. If p is a strong authority or hub for some reasonably strong cluster, then the other authorities or hubs in this cluster are considered similar to p.

Note that the only difference from the basic experiment in Section 3 is in the construction of the set S: rather than choose it based on the presence of a query string, we choose it based on adjacency to p in the underlying graph.

If p is not a sufficiently strong hub or authority then one cannot really interpret the resulting pages as being similar to p. However, the resulting output still gives sets of hubs and authorities for the region of the WWW in the "neighborhood" of p. This could be of value in performing a type of broad-topic classification of p.

For pages that locally are strong authorities, a number of fairly striking results have been produced. The following two lists of authorities are with respect to the home pages of Honda Motor Company and the New York Stock Exchange respectively.

```
(www.honda.com) Authorities 0
0) 0.202331 http://www.toyota.com/
        Welcome to @Toyota
1) 0.199263 http://www.honda.com/
```

```
        Honda
2) 0.192716 http://www.ford.com/
        Ford Motor Company
3) 0.173613 http://www.bmwusa.com/
        BMW of North America, Inc.
4) 0.162200 http://www.volvocars.com/
        VOLVO
5) 0.158151 http://www.saturncars.com/
        Welcome to the Saturn Web Site
6) 0.155788 http://www.nissanmotors.com/
        NISSAN - ENJOY THE RIDE
7) 0.145088 http://www.audi.com/
        Audi Homepage
8) 0.139752 http://www.4adodge.com/
        1997 Dodge Site
9) 0.136581 http://www.chryslercars.com/
        Welcome to Chrysler

(www.nyse.com) Authorities 0
0) 0.208114 http://www.amex.com/
        The American Stock Exchange - The Smarter Place to Be
1) 0.146286 http://www.nyse.com/
        New York Stock Exchange Home Page
2) 0.134995 http://www.liffe.com/
        Welcome to LIFFE
3) 0.129817 http://www.cme.com/
        Futures and Options at the Chicago Mercantile Exchange
4) 0.120296 http://update.wsj.com/
        The Wall Street Journal Interactive Edition
5) 0.118149 http://www.nasdaq.com/
        The Nasdaq Stock Market Home Page - Reload Often
6) 0.117228 http://www.cboe.com/
        CBOE - The ChicagoBoard Options Exchange
7) 0.116350 http://www.quote.com/
        1- Quote.com - Stock Quotes, Business News, Financial Market
8) 0.113538 http://networth.galt.com/
        NETworth
9) 0.109859 http://www.lombard.com/
        Lombard Home Page
```

A frequent occurrence is that a very strong cluster of home pages for computer companies or search engines appears in the output, regardless of the starting page. The reason for this is clear — on the WWW, there is a high density of links to such pages in the neighborhood of nearly *every* page. Via the clustering method of Section 2.3, one hopes to produce a strong non-principal eigenvector which places the cluster of computer companies at one extreme,

and the set of pages one is truly interested in at the other. The following example — the home page for the New York Times — gives an indication of this phenomenon.

```
(www.nytimes.com) Authorities 0
0) 0.287919 http://www.yahoo.com/
        Yahoo!
1) 0.181236 http://www.nytimes.com/
        The New York Times on the Web
2) 0.170931 http://www.usatoday.com/
        USA TODAY
3) 0.165114 http://www.cnn.com/
        CNN Interactive
4) 0.124062 http://www.mckinley.com/
        Welcome to Magellan!
5) 0.120629 http://www.altavista.digital.com/
        AltaVista Search:  Main Page
6) 0.119957 http://www.excite.com/
        Excite
7) 0.117229 http://www.microsoft.com/
        Welcome to Microsoft
8) 0.108421 http://www.whitehouse.gov/
        Welcome to the White House
9) 0.107821 http://www.lycos.com/
        Lycos, Inc.  Home Page
```

Thus, the principal eigenvector produces essentially a mixture of computer/Internet companies and news organizations. The first non-principal eigenvector separates this mixture into two distinct clusters.

```
(www.nytimes.com) Authorities +1
0) 0.111950 http://www.microsoft.com/
        Welcome to Microsoft
1) 0.110140 http://www.ibm.com/
        IBM Corporation
2) 0.101674 http://www.apple.com/
        Apple Computer
3) 0.100671 http://www.hp.com/
        Welcome to Hewlett-Packard
4) 0.098997 http://www.sun.com/
        Sun Microsystems
5) 0.097509 http://www.intel.com/
        Welcome to Intel
6) 0.097075 http://www.novell.com/
        Novell World Wide:  Corporate Home Page
7) 0.087379 http://www.ustreas.gov/
        Welcome To The Department of Treasury
```

```
8) 0.084578 http://www.compuserve.com/
        Welcome to CompuServe
9) 0.081480 http://www.lcs.mit.edu/
        MIT Lab for Computer Science Web Page

(www.nytimes.com) Authorities -1
0) -0.220348 http://www.nytimes.com/
        The New York Times on the Web
1) -0.169549 http://www.usatoday.com/
        USA TODAY
2) -0.138866 http://www.cnn.com/
        CNN Interactive
3) -0.091273 http://www.sjmercury.com/
        Mercury Center
4) -0.080803 http://www.chicago.tribune.com/
        The Chicago Tribune
5) -0.076501 http://www.washingtonpost.com/
        Welcome to WashingtonPost.com
6) -0.074173 http://www.cbs.com/
        EYE ON THE NET @ CBS
7) -0.066685 http://www.npr.org/
        Welcome to NPR
8) -0.063137 http://www.telegraph.co.uk/
        Electronic Telegraph
9) -0.061040 http://nytimesfax.com/
        TimesFax
```

4.2 Intranet Analysis

Finally, we arrive at what is perhaps the simplest experiment of all: given the hyperlinked set of all pages T residing on a given server, we run the algorithm on this set T. Note that in this setting, we can only consider links that have both endpoints on the server. One point that must be resolved in performing this experiment is how to re-define the notion of a *transverse link* — since all pages now have the same domain name, there would not be *any* transverse links under the old definition. We tried both of the following options:

(i) Define all links to be transverse links.

(ii) Say that a link is *intrinsic* if one of its endpoints is a descendant of the other in the URL tree. A link is transverse iff it is not intrinsic.

We ran the algorithm on the domain `www.research.ibm.com`, which consists of approximately 10,000 nodes and 20,000 intra-site links. Fewer than 5% of the links are intrinsic; hence it is not surprising that the results were not substantially affected by the decision between options (i) and (ii) above. (Option (ii) does have the effect of turning the root page, which is otherwise a strong authority, into an isolated node.)

A basic difference between the results of this experiment and those of the previous experiments is that in the present case, no strong hub pages emerged. One reason for this is likely the scarcity of personal home pages on the server. But at a more basic level, one also has the problem that hub pages typically reference a geographically widespread collection of sites, and hence are difficult to detect if one can only consider links that stay within a single domain. One might argue that for intranets of this type, a metaphor different from the hub/authority relationship is needed in order to extract information from the link structure; this would be an interesting question to consider further.

However, one validation of our clustering algorithm lies in the following: even when intrinsic links were omitted, the clusters produced by our algorithm corresponded quite closely to subtrees of the URL hierarchy.

5 Diffusion and Lexical Scores

Finally, we turn to a discussion of some of the typical ways in which the algorithm fails, and some possible extensions of the method to make it more robust against the most common difficulties. We will be focusing on the query-based experiments of Section 3, in which a root set S is generated by a search engine, a set T is grown around it, and the algorithm is run on T. The basic phenomenon we investigate here is what one could call *diffusion*: the algorithm converges to a set of hubs and authorities that are not focused on the original topic. This has a greater tendency to occur as the query topic becomes more focused; the reasons why this should be the case have already been discussed to some extent in Section 3. Analogously, it has a greater tendency to occur when the query topic is "close" to another topic with much greater representation on the WWW.

5.1 Basic Examples

As a good first example, we consider the the experiment from Section 3 on the query ("medical conferences") (again from the YAHOO *Health/Medicine* list). Although AltaVista indexes roughly 600 pages containing the term, the algorithm on the one-step neighborhood set T essentially converges to authoritative pages for the topic "medicine" in general:

```
("medical conferences") Authorities 0
0) 0.087812 http://www.cdc.gov/
        Centers for Disease Control and Prevention Home Page
1) 0.083831 http://www.ohsu.edu/cliniweb/
        Cliniweb
2) 0.081350 http://www.bmj.com/bmj/
        BMJ
```

It is not surprising that this should have happened: pages on medical conferences typically link to a large number of general medical resources, and hence these acquire a lot of authority. More interesting, perhaps, is to ask why a similar phenomenon did not occur for the ten medical topics considered in Section 3.4. Although the issue is clearly quite subtle, one can observe that (a) the ten earlier topics are considerably "larger" than the current one, in that

AltaVista indexes roughly 20,000 pages for most of them; and (b) there are correspondingly more resource pages focused on these topics, which can "freeze" the authority weight at more specific pages and prevent it from diffusing to more general ones.

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

```
("WWW conferences") Authorities 0
0) 0.088844 http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/whats-new.html
        The What's New Archive
1) 0.088839 http://www.w3.org/hypertext/DataSources/WWW/Servers.html
        World-Wide Web Servers:  Summary
2) 0.087786 http://www.w3.org/hypertext/DataSources/bySubject/Overview.html
        The World-Wide Web Virtual Library:  Subject Catalogue
```

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

5.2 A Lexical Scoring Function

There is a range of techniques one could try implementing to help prevent diffusion; unfortunately, many of them strongly interfere with the positive features of the algorithm as it stands. Specifically, it is tempting to weight each page by a lexical score derived from the query string, and incorporate these weights into the iterations of the algorithm as it computes hubs and authorities. However, while this would undoubtedly keep the underlying topic more strongly in focus, we maintain that it would damage the algorithm's ability to perform the basic task of locating authoritative sources. In particular, going back to the initial examples in Section 3, consider that `www.microsoft.com` does not contain the term "Gates," `www.eff.org` does not contain the term "censorship," and Netscape's main pages do not contain the term "browser." It is easy to generate many other natural examples of this phenomenon.

Thus, it is a *feature* of our algorithm that it was able to ignore the absence of the query term from these pages, and still give them large authority weight. The point, of course, is that a large number of other pages simultaneously used the term and pointed to these pages.

We propose the following approach, which allows the clusters to develop as before, and only re-introduces the query term once they have been constructed. We first run the basic algorithm on the set T, producing some relatively large number k of clusters. (We noted earlier that only a relatively small number of clusters are associated with eigenvalues of non-trivial magnitude, and we clearly wish to restrict ourselves to these.) We then apply a lexical scoring function to each cluster as a whole, and rank the clusters according to this function. Of course, any attempt to use a lexical scoring function will suffer to a greater or

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

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

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

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

(iii) The score for the cluster C is then

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

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

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

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

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

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

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

```
("WWW conferences") Authorities -11
0) -0.097089 http://www.igd.fhg.de/www95.html
        Third International World-Wide Web Conference
1) -0.091676 http://www.csu.edu.au/special/conference/WWWWW.html
        AUUG'95 and Asia-Pacific WWW'95 Conference and Exhibition
2) -0.090432 http://www.ncsa.uiuc.edu/SDG/IT94/IT94Info.html
        The Second International WWW Conference '94:  Mosaic and the
3) -0.083832 http://www.w3.org/hypertext/Conferences/WWW4/
        Fourth International World Wide Web Conference
4) -0.079349 http://www.igd.fhg.de/www/www95/papers/
        WWW'95:  Papers
```

5.3 Term Mixtures

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

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

In order to make these notions more concrete, let us consider the following example from the YAHOO *Entertainment/Movies* list: the query (+movies +awards). (The syntax here indicates that both words must appear in the pages returned by the search engine.) Similarly to the previous examples, the primary cluster diffused to the more general topic "movies":

```
(+movies +awards) Authorities 0
0) 0.291012 http://www.disney.com/
        Disney.com Home Page - Welcome
1) 0.278708 http://www.hollywood.com/
        Hollywood Online
2) 0.217913 http://www.paramount.com/
        Paramount Pictures Online
```

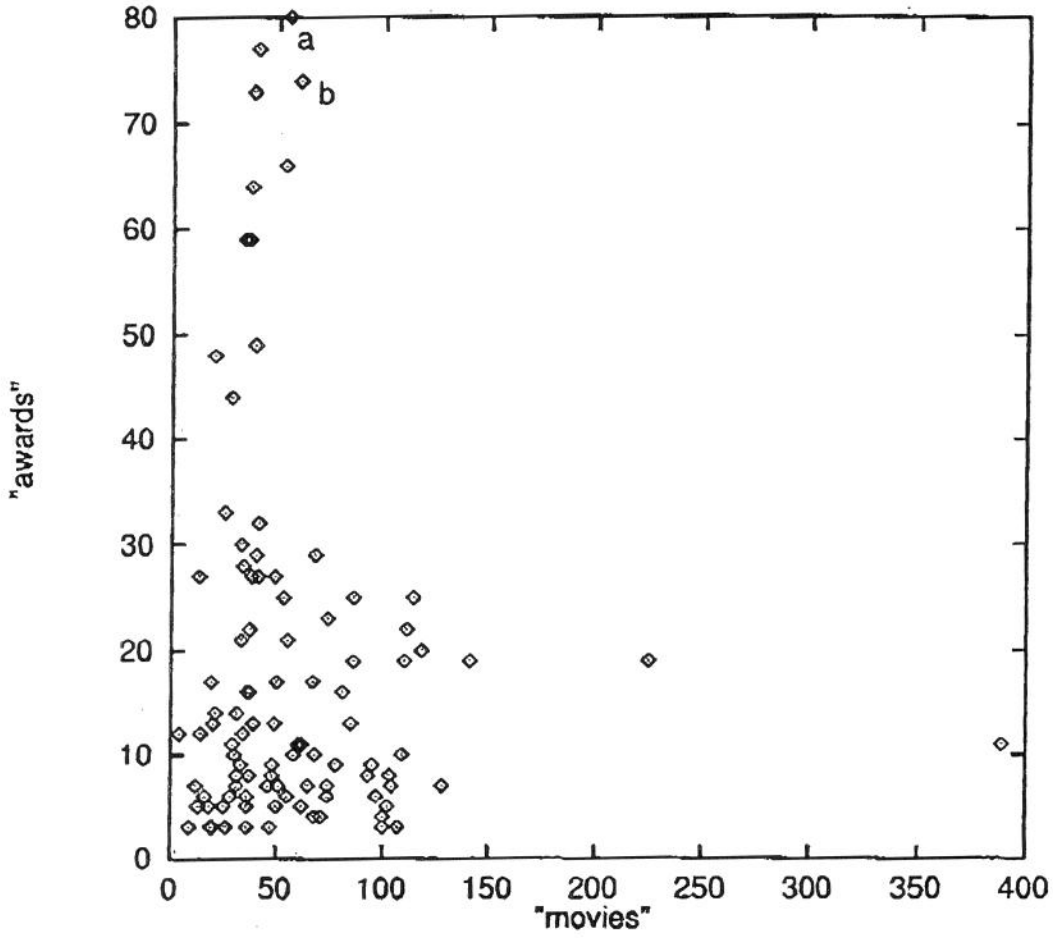

Figure 4: Term occurrences in (+movies +awards)

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

For each of the 99 clusters examined, our scoring method produced a two-coordinate vector (one coordinate for each of "movies" and "awards"). In Figure 4, we provide a scatter-plot of the resulting 99 points in the plane. Seven of the 99 points are maximal. It is also interesting to consider the *positive hull* of the point set: by this we mean the set of all points that lie on a line of negative slope which does not separate the point set. Only two of the points lie on the positive hull.

The two maximal points labeled "a" and "b" in Figure 4 turned out to be arguably the most relevant clusters. For the one labeled "a", the top five hub pages include the YAHOO and Zia pages for *Movies/Academy Awards*; the top five authorities are as follows:

```
(+movies +awards) Authorities -37
0) -0.118528 http://www.ampas.org/
        The Academy of Motion Picture Arts and Sciences
```

```
1) -0.110994 http://www.mnet.fr/dian.ying/
        Index dian ying
2) -0.108717 http://ddv.com/Oscarnet/
3) -0.108453 http://oscars.guide.com/
        THE ENVELOPE PLEASE Interactive Guide to Academy Awards & Oscars
4) -0.101703 http://www.hype.com/movies/oscars/home.htm
        You predict the Oscars for the 68th Annual Academy Awards!
```

For the cluster labeled "b", the top five hub pages include the YAHOO and Zia pages for *Movies/Awards*; the top five authorities are as follows:

```
(+movies +awards) Authorities +37
0) 0.223290 http://www.bafta.org/
        Croeso i Bafta Cymru - Welcome to Bafta Cymru
1) 0.211654 http://www.choiceawards.com/
2) 0.211654 http://www.sunflower.org/~henryj/movie.htm
3) 0.211654 http://www.razzies.com
        The Golden Raspberry Award Foundation (The "Razzies")
4) 0.211654 http://www.emerson.edu/acadepts/mc/EVVY_HP.HTML
```

Let us return to the issue of ordering the cluster scores. Since our purpose is only to cover some of the most basic possibilities, our discussion here will be brief. Recall that each cluster C has an m-coordinate score $a_C = (\alpha_{s_1}(C), \ldots, \alpha_{s_m}(C))$. One natural method would be to sum all the coordinate values in a_C. However, this seems not to be a robust approach, for the reason that certain terms appear to exhibit much greater variance than others in the number of times they occur. Such terms could then wield too great an influence in the scoring function. (Considering the plot in Figure 4, the right-most point would be ranked highest in this measure on the basis of its x-coordinate alone.)

One simple proposal that we feel is borne out better, both intuitively and in our experiments, is the following. For a vector a, define $\mu(a)$ to be the minimum of its coordinate values, and rank the clusters according the values of $\mu(a_C)$. In this way, one tries explicitly to find clusters in which *all* query terms occur as much as possible. For the above example, the μ-optimal cluster is the one labeled "b." In Figure 5, we show the results of ranking based on μ for a subset of the topics from the YAHOO *Movies/Genres* list. The 39 clusters associated with the first 20 eigenvectors were considered; the second column gives the number of maximal clusters, and the third column gives the number lying on the positive hull. Overall, the lexical scoring approach discovered relevant clusters associated with weaker eigenvectors in several of the cases.

Ultimately, it seems clear that the problem of scoring and ranking the clusters produced by our algorithm, based on association with a query term, is quite a wide-open issue.

6 Conclusion

We feel that there are a number of directions arising in this work that would be interesting to pursue further. Several of these have been noted at points in the paper; here we summarize

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

Figure 5: Optimal clusters for *Entertainment/Movies/Genres* under total term measure.

what we believe to be three of the most substantial ones.

(i) We observed at the outset — and the description of the method should make this clear — that the approach we describe need not be restricted to hypermedia. At the most basic level, one can investigate its application to the cross-referencing structure of collections of scientific papers or patents. But more generally, there are a number of naturally arising directed graph structures in which one can find clear interpretations for the notions of "hubs" and "authorities." Consider, for example, the implicit analogy drawn in [25] between the relationships among modules in a large software system and the basic measures used in bibliometrics. One can also consider the use of our method on graphs defined by financial transactions, or communications (e.g. e-mail), among a large set of individuals and organizations.

(ii) In the query-based experiments of Section 3, we used a very basic technique for constructing the "one-step neighborhood" T. It is natural to consider more sophisticated methods of defining a neighborhood for the root set provided by the search engine. One promising notion is that of adaptively updating the set as the algorithm proceeds. One could, for example, consider running multiple successive phases of the algorithm — at the end of each phase, one attempts to identify a good set of nodes to use as a root set in the next phase. Here, the notion of "a good set of nodes" is clearly what contains all the complexity. Presumably this judgment would involve both the hub and authority weights, and — to maintain relevance to the initial query — some re-introduction of the textual content of the pages. This, indeed, brings us to our third point.

(iii) One of the striking features of the query-based version of the algorithm is the frequency with which it remains focused on the initial query topic, when this topic is sufficiently "broad." More work, both analytical and experimental, needs to be done in order to better understand the boundary between the set of queries on which the algorithm remains focused, and the set on which it will *diffuse* (typically to a more general topic). And for those on which the pure algorithm does not succeed, we are interested in determining the best way to incorporate the textual content of the pages. Our experiments in Section 5 indicate some of the basic approaches that are possible in this direction; but there is clearly a range of

further techniques that could be investigated.

Acknowledgements

I thank Prabhakar Raghavan for invaluable on-going discussions on aspects of this work, and collaboration on several of the experiments presented here; Robert Kleinberg for generously sharing, as always, his insights on these problems; Rob Barrett for suggesting the intranet experiment and providing me with the initial data; and Tryg Ager, Soumen Chakrabarti, Alan Hoffman, Nimrod Megiddo, Sridhar Rajagopalan, Eli Upfal, and many others within IBM Almaden and Watson for their valuable comments and suggestions.

References

[1] G.O. Arocena, A.O. Mendelzon, G.A. Mihaila, "Applications of a Web query language," *Proc. 6th International World Wide Web Conference*, 1997.

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

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

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

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

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

[7] W.E. Donath, A.J. Hoffman, "Algorithms for partitioning of graphs and computer logic based on eigenvectors of connections matrices," *IBM Technical Disclosure Bulletin*, 15(1972), pp. 938–944.

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

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

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

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

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

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

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

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

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

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

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

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

[20] 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.

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

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

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

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

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

[26] 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.

[27] 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.

[28] 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.

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

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

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

[32] 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.

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

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

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