Applications of linear algebra in information retrieval and hypertext analysis

Jon Kleinberg * Andrew Tomkins†

1 Overview

Information retrieval is concerned with representing content in a form that can be easily accessed by users with information needs [61, 65]. A definition at this level of generality applies equally well to any index-based retrieval system or database application; so let us focus the topic a little more carefully. Information retrieval, as a field, works primarily with highly unstructured content, such as text documents written in natural language; it deals with information needs that are generally not formulated according to precise specifications; and its criteria for success are based in large part on the demands of a diverse set of human users.

Our purpose in this short article is not to provide a survey of the field of information retrieval — for this we refer the reader to texts and surveys such as [25, 29, 51, 60, 61, 62, 63, 65, 70]. Rather, we wish to discuss some specific applications of techniques from linear algebra to information retrieval and hypertext analysis. In particular, we focus on spectral methods — the use of eigenvectors and singular vectors of matrices — and their role in these areas.

After briefly introducing the use of vector-space models in information retrieval [52, 65], we describe the application of the singular value decomposition to dimension-reduction, through the Latent Semantic Indexing technique [14]. We contrast this with several other approaches to clustering and dimension-reduction based on vector-space models.

*Department of Computer Science, Cornell University, Ithaca NY 14853. Email: kleinber@cs.cornell.edu. Supported in part by an Alfred P. Sloan Research Fellowship and by NSF Faculty Early Career Development Award CCR-9701399.

†IBM Almaden Research Center, San Jose CA 95120. Email: tomkins@almaden.ibm.com.

We then turn to hyperlinked corpora — collections of documents with an underlying link structure. The emergence of the World Wide Web [2] has led to a surge of interest in the problem of information retrieval in such domains; we describe some approaches that apply spectral methods to link structures for information discovery tasks (e.g. [8, 43]). There are connections between this work and earlier work in sociology and citation analysis [24], and we discuss this as well.

2 Information retrieval and the vector space model

The language of linear algebra made its appearance quite early in information retrieval, through the use of *vector-space models* [52, 65]; these models have formed the basis for information retrieval frameworks such as Salton's SMART system (see e.g. [10, 65]). We begin with a set of d *documents* and a set of t *terms*. We model each document as a vector x in the t-dimensional space $\mathbf{R}^t$ — it has one coordinate for each term. The j^{th} coordinate of x is a number that measures the association of the j^{th} term with respect to the given document — it is generally defined to be 0 if the document does not contain the term, and non-zero otherwise.

The problem of how to define the non-zero entries in such a vector is known as *term-weighting*, and it has been the subject of a large amount of work; see e.g. [17, 64, 65, 68]. Perhaps the simplest formulation is to set $x_j = 1$ if the j^{th} term occurs at all in the document. More general approaches based on *term frequency* and *inverse document frequency* take into account the number of times the term occurs in the document, and the total number of documents in the corpus in which the term occurs.

The representation of documents by vectors in Euclidean space allows one to bring geometric methods to bear in analyzing them. At the simplest level, the representation naturally suggests numerical similarity metrics for documents, based on the Euclidean distance or the inner product. Again, many related metrics have been proposed, and we discuss one representative — the *cosine measure* [65] — that will help motivate some of

the developments to follow. Let x and y be two document vectors. We define their *cosine similarity* by the equation

$$\mathrm{sim}(x, y) = \frac{x \cdot y}{|x||y|},$$

where the inner product $x \cdot y$ is the standard vector dot product, defined as $\sum_{i=1}^{t} x_i y_i$, and the norm in the denominator is defined as $|x| = \sqrt{x \cdot x}$. We term this the cosine similarity because for any two unit vectors, it is simply the cosine of the angle between them.

Numerical similarity metrics on documents suggest natural approaches for similarity-based indexing (e.g. [66]) — by representing textual queries as vectors and searching for their *nearest neighbors* in a collection of documents — as well as for clustering (e.g. [40]). Of course, in any application with a large number of underlying terms, these vector operations are being carried out in a huge number of dimensions. Very high dimensionality can be a problem not only from the point of view of computational efficiency, but also because the large number of terms leads to sets of vectors with very sparse patterns of non-zeroes, in which relationships among terms (e.g. synonymy) can be difficult to detect or exploit. An effective method for reducing the dimension of the set of vectors, without seriously distorting their metric structure, offers the possibility of alleviating both these problems. We turn to this issue next, beginning with some fundamental background material from linear algebra.

3 Linear algebra, eigenvalues, and the singular value decomposition

Given a set of d vectors representing a collection of d documents, we can construct a $t \times d$ matrix in which each document vector constitutes one of the columns. Our interest will be in transforming this matrix to one that has low rank; this will correspond to a dimension-reduction of the set of documents. To make this notion precise, we introduce the following definitions; we refer the reader to linear algebra texts such as [38, 69] for further details.

First, let M be an $n \times n$ matrix with real numbers as entries. An *eigenvalue* of M is a number λ with the property that, for some vector ω, we have $M\omega = \lambda\omega$. Such a vector is called an *eigenvector* associated with λ. The set of all eigenvectors associated with a given λ is a subspace of $\mathbf{R}^n$, and the dimension of this space will be referred to as the *multiplicity* of λ. If M is a symmetric matrix — one of the main cases of interest for our purposes — then 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 $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$, listing each a number of times equal to its multiplicity. For symmetric M, we can choose an eigenvector $\omega_i(M)$ associated with each $\lambda_i(M)$ so that the set of vectors $\{\omega_i(M)\}$ forms an orthonormal basis of $\mathbf{R}^n$ — each is a unit vector, and each pair of them is orthogonal.

We say that a matrix Q is *orthogonal* if $Q^T Q = I$, where Q^T denotes the *transpose* of the matrix M, and I represents the *identity matrix* — a diagonal matrix with all diagonal entries equal to 1. If M is a symmetric $n \times n$ matrix, Λ is the diagonal matrix with diagonal entries $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$, and Q is the matrix with columns equal to $\omega_1(M), \ldots, \omega_n(M)$, then it is easy to verify that Q is an orthogonal matrix and $Q\Lambda Q^T = M$.

Thus, the eigenvalues and eigenvectors provide a useful "normal form" representation for symmetric square matrices in terms of orthogonal and diagonal matrices. In fact, there is a way to extend this type of normal form to matrices that are neither symmetric nor square, as we now discuss.

Theorem 1 (Singular Value Decomposition (SVD)) *Every $m \times n$ matrix A can be written $A = U\Sigma V^T$ where U and V are orthogonal, and Σ is diagonal.*

That is, we can rewrite any matrix A as follows:

$$\overbrace{\begin{pmatrix} \\ A \\ \\ \end{pmatrix}}^{m \times n} = \overbrace{\begin{pmatrix} \\ U \\ \\ \end{pmatrix}}^{m \times m} \cdot \overbrace{\begin{pmatrix} \\ \Sigma \\ \\ \end{pmatrix}}^{m \times n} \cdot \overbrace{\begin{pmatrix} \\ V^T \\ \\ \end{pmatrix}}^{n \times n}$$

We refer to the diagonal entries of Σ as the *singular values* of A. Using the SVD directly, we can write $A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^2 V^T$. Likewise, $AA^T = U\Sigma^2 U^T$. It follows that the columns of U and V represent the eigenvectors of AA^T and $A^T A$ respectively, and the diagonal entries of Σ^2 represent their (common) set of eigenvalues.

What does the SVD have to do with dimension-reduction? To begin with, we notice the following fact. Suppose we build a matrix by keeping only the k largest singular values, and multiplying them by the appropriate rows and columns of U and V: $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, where u_i is the i^{th} column of U and v_i is i^{th} column of V. We first notice that the matrix A_k has rank at most k. Moreover, it is a reasonable rank-k approximation to the original matrix A in the sense that we have only "zeroed out" the small singular vectors of A. Fuzzy as this intuitive description is, it can be converted to a surprisingly concrete statement: the matrix A_k described above is the best rank-k approximation to A in the matrix 2-norm.

Theorem 2 (Eckart and Young; see [38]) *Let the SVD of A be given by $U\Sigma V^T$, let $r = \mathrm{rank}(A)$, and $k < r$. If*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

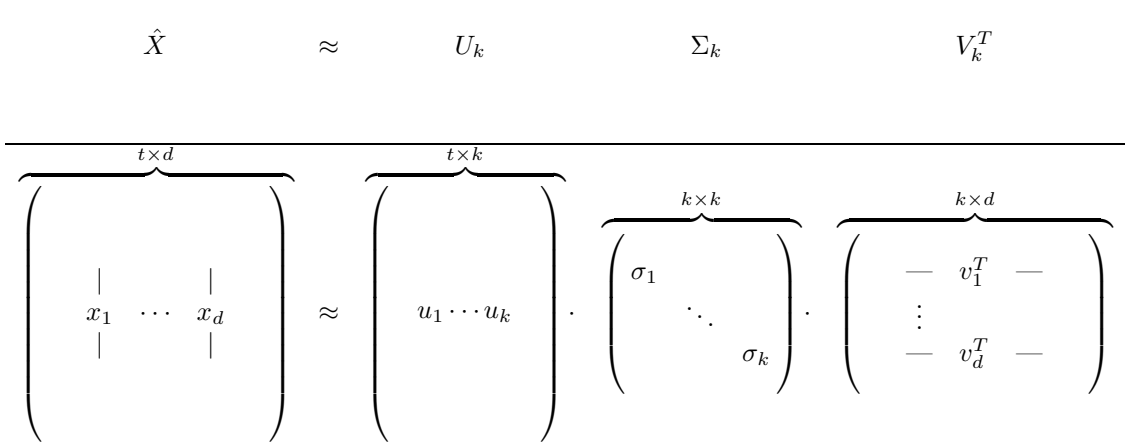

Figure 1: The LSI method

then

$$\min_{\text{rank}(B)=k} ||A - B||_2 = ||A - A_k||_2 = \sigma_{k+1}.$$

4 Latent semantic indexing

We use the machinery of the previous section to outline the technique of *latent semantic indexing* (LSI), a powerful approach to dimension-reduction in information retrieval developed by Deerwester et al. [14].

Suppose we have a collection of d document vectors, over a set of t terms, and we construct the matrix X whose columns consist of these vectors. We can view Theorem 2 as having the following interpretation in this setting (see Figure 1). We fix a value k, consisting of a (relatively small) number of dimensions in which we want to represent the documents. By retaining only the k largest singular values in the SVD of X, we obtain a $k \times k$ diagonal matrix Σ_k with these singular values as its diagonal entries, and we write $\hat{X} = U_k \Sigma_k V_k^T$, with U_k a $t \times k$ matrix and V_k^T a $k \times d$ matrix. We can view each of the d columns of the matrix $\Sigma_k V_k^T$ as representing one of the documents; in this representation, the documents have been projected into the k-dimensional subspace spanned by the columns of the matrix U_k.

The choice of k is a non-trivial issue — there is a trade-off between the amount of dimension-reduction and the accuracy of the resulting document representation. In many applications, k is on the order of several hundred, while the original representation involves a dimension in the thousands or tens of thousands.

An effect of our projection is that each term no longer occupies a distinct dimension; rather, each of the k new dimensions corresponds to a vector — a column of U_k — that is a weighted sum of terms. It has been found (e.g. [14]) that terms which co-occur in the corpus (e.g. "car," "automobile," and "vehicle") have similar weights in these vectors: the low-dimensional projection has reduced the "noise" introduced by the effects of co-occurring terms on our similarity measure. This has consequences for several basic applications employing, for example, the cosine measure.

1. We can compute document-document similarities using the matrix V_k. Given a document in which we know a user is interested, we can suggest other documents that might also be of interest, even if they do not use the same terms.

2. We can compute term-term similarities using matrix U_k. Given a one-word query, for instance, we can expand it to include words that tend to co-occur in the corpus, augmenting "car" with "automobile" and "vehicle."

3. We can compute term-document similarities directly using the entries of $\hat{X}$ — for instance, when the user enters a search term, we can return the documents that are most similar to the search term. Analogously to the previous two cases, the top-ranked document may not necessarily contain the search term.

A number of papers have tried to develop interpretations of the columns of U_k, the axes of the reduced space, in terms of the basic information retrieval applications. Since, as we observed above, each can be viewed as a weighted combination of terms, some previous work has even offered the view that they represent the fundamental "concepts" that underlie the collection of documents [47].

Some evaluations of the effectiveness of LSI are given in [18, 19, 20]; computational issues are considered in [3]; and additional perspectives and extensions related to LSI can be found in [4, 5, 6, 21, 32, 33, 34, 48]. LSI has been applied in a variety of domains, including cross-language retrieval [22, 23, 49] and information filtering [20, 28].

Papadimitriou et al. [56] recently studied LSI at an analytical level, by employing a probabilistic model of term usage across a "clustered" collection of documents. Essentially, their model consists of k different *topics* (hidden from the retrieval algorithm); a document on a given topic τ is generated by a repeated random selection of terms from a probability distribution F_τ over terms. For different topics τ and τ', there is a technical condition enforcing that the distributions F_τ and $F_{\tau'}$ are "well-separated." This is the sense in which the collection of documents is "clustered" — each document belongs to a topic that induces a distinctive distribution over terms. The main result of [56] is that, when the distributions induced by different topics are sufficiently separated, the k-dimensional subspace produced by LSI yields, with high probability, sharply defined clusters among documents of the k different topics with respect to the cosine measure. This provides a concrete analytical sense in which LSI in certain settings is able to uncover semantically "meaningful" associations among documents with similar patterns of term usage, even when they do not actually use the same terms.

Other Approaches to Dimension-Reduction. Papadimitriou et al. also observed that, if the reduced dimension k we are seeking is sufficiently large relative to the number of documents, it is not necessary to use the full power of the singular value decomposition — a *random* projection is sufficient. This is the content of the following result of Johnson and Lindenstrauss [41], sharpened by Frankl and Maehara [30].

Theorem 3 *For $\varepsilon \in (0, \frac{1}{2})$ and any positive integer n, let*

$$k(n,\varepsilon) = \lceil 9(\varepsilon^2 - 2\varepsilon^3/3)^{-1} \log n \rceil + 1 = O\left(\frac{\log n}{\varepsilon^2}\right).$$

If $n > k(n,\varepsilon)^2$ then for any n-point set S in R^n, there exists a map $f : S \to R^{k(n,\varepsilon)}$ such that for all $u, v \in S$,

$$(1-\varepsilon)|u-v|^2 < |f(u)-f(v)|^2 < (1+\varepsilon)|u-v|^2.$$

The map f in the statement of the theorem can be constructed very easily: one chooses a random subspace of dimension $k(n,\varepsilon)$ and applies an orthogonal projection into this subspace, followed by a uniform rescaling. The desired approximation property then holds with high probability. Papadimitriou et al. suggested that the SVD calculations in Latent Semantic Indexing could be sped up by first projecting into a random subspace of dimension $O(\varepsilon^{-2} \log n)$, and then computing the SVD [56]. Further approaches to produce highly efficient approximation algorithms for the SVD through random sampling appear in Frieze et al. [31] and Drineas et al. [16].

In a different direction, Baker and McCallum [1] applied the technique of *distributional clustering* [57] to the task of dimension-reduction. Distributional clustering is a framework that also seeks to cluster terms based on co-occurrence; as opposed to LSI, however, it is based on an information-theoretic model that represents terms via the probability distributions they induce over features with which they co-occur. Experiments in [1] showed that this technique compared favorably with LSI, as well as with several other techniques [9, 45, 72], for reducing feature dimensionality in a document classification task.

For other recent approaches to dimension-reduction, and its relation to the general area of *feature selection*, see [44, 45, 71, 72].

5 Hyperlinked domains and spectral graph theory

In the remainder of this article, we move from the pure framework of terms and documents to a setting in which documents are connected by an underlying link structure. This captures the problem of information retrieval on the World Wide Web, where one can make use of both the textual content of documents as well as the patterns of linkage among them. The study of links as a means of understanding the informational content of a collection of documents predates the Web and other modern hypertext systems, however; the implicit linkage defined by citations among scholarly papers has been a fundamental object of study in the field of *bibliometrics*, or citation analysis [24].

Dimension-reduction and clustering based on the singular value decomposition has been applied to link structures in the field of bibliometrics. A natural way to obtain a matrix from an underlying citation structure is by the following construction: given a collection of n documents, we define an $n \times n$ matrix A for which the (i,j) entry is equal to 1 if document i cites document j, and 0 otherwise. One can recognize this as the *adjacency matrix* of the directed graph whose directed edges correspond to the citations among documents.

By analogy with the vector-space model in information retrieval, one could use the rows and columns of this matrix A to represent the documents in the citation structure. The application of the singular value decomposition to such vector representations has been investigated by Small [67], McCain [54], and Noma [55]. More recently, the application of dimension-reduction techniques to such vector representations of WWW pages has been employed by Larson [50] and by Pitkow and Pirolli [59].

The use of eigenvectors for clustering link structures is a technique that can be understood directly at the level of the underlying graph model. Indeed, this type of eigenvector-based clustering is a topic that has received considerable study in the area of graph algorithms, beginning with the foundational work of Donath and Hoffman [15] and Fielder [27] on *spectral partitioning* heuris-

tics. A large body of results now exists, relating spectral properties of adjacency matrices to combinatorial properties of the associated graphs; see the book by Chung [13] for a recent overview.

6 Impact, influence, and authority in linked domains

Let us now turn to an issue different from the problems of clustering and dimension-reduction, which also constitutes a central question in the development of information retrieval techniques. Suppose we have access to a large set of documents relevant to a given topic, and we wish to automatically select the most "important" ones. One could imagine many settings in which this notion arises: we may be surveying the scientific literature, looking for "seminal" papers on quantum mechanics; or we may be searching the WWW, looking for the most "authoritative" pages on cryptography. Note the difference between this issue and what we have been discussing previously — we are not concerned here with representing the set of *all* relevant material, but rather with the problem of *filtering*, from a large volume of relevant content, a small set of the most significant documents.

Links provide a natural mechanism for quantifying notions of "importance"; in both scientific citations and hypertext, a link can indicate the judgment of the author of one document as to the importance of another document. Indeed, the use of links to measure "social standing" has also been investigated in the field of social networks; link-based measures of standing have been proposed in a sociometric context by Katz [42], Hubbell [39], and others. For purposes of the present discussion, however, we will focus on the use of links in the areas of citation analysis and hypertext.

Citation Analysis. The most widely-used measure of "importance" in citation analysis is Garfield's *impact factor* [35], used to provide a numerical assessment of journals in the 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 [24]. Thus, the impact factor is based fundamentally on a pure counting of the number of links pointing into each node of the citation network.

Pinski and Narin [58] proposed a significant variation on this notion, based on 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. The concrete construction of Pinski and Narin, as modified by Geller [36], is the following. The measure of standing of journal j will be called its *influence weight* and denoted w_j. Given a set of n journals, one constructs an $n \times n$ matrix A in which the (i, j) entry specifies the "connection strength" from i to j: it is 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 $w_j = \sum_i A_{ij} w_i$; in other words, if w is the vector whose j^{th} entry is w_j, then we have $A^T w = w$. Thus, the set of influence weights under the Pinski-Narin definition is precisely an eigenvector of the matrix A^T associated with the eigenvalue 1. Geller [36] observed that the influence weights also correspond to a 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. Indeed, one can verify that a stationary set of probabilities w for such a process must satisfy the equation $A^T w = w$. (See e.g. the text by Feller [26].)

WWW Search. In the setting of the World Wide Web, Brin and Page proposed a method for ranking the "importance" of Web pages [8], based on a model of a "random browser." 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 collection of pages. The stationary probability of node j in this random process will correspond to the "rank" of j, referred to as its *page-rank*. Note that the random jump is crucial to prevent the random process from getting stuck in "dead-ends" in the link structure of the Web. If we consider the matrix B whose (i, j) entry is the probability of going directly from page i to page j in this process, and let r be the vector whose j^{th} coordinate is the page-rank of page j, then we are seeking a solution to the equation $B^T r = r$; so we are seeking an eigenvector of B associated with the eigenvalue 1.

Kleinberg proposed a different model for the conferral of authority on the WWW [43]. He argued that in many settings on the Web, prominent authorities do not "endorse" one another directly — consider, for example, a collection of prominent corporate home pages in a common area. Rather, for many broad topics, authority is conferred on thematically related, prominent pages by a set of potentially unrecognized *hub pages*, which have a large number of links to many relevant authorities. Thus, hubs and authorities exhibit a mutually reinforcing relationship: a good authority is a page that is pointed to by many good hubs, while a good hub is page that points to many good authorities. Numerically, one can assign a *hub weight* h_j and an *authority weight* a_j to each page j. If we let h and a denote the

normalized vectors whose coordinates correspond to the hub and authority weights respectively, and we let A denote the adjacency matrix of the link graph as defined earlier, one could construct estimates for the "ideal" hub and authority weights iteratively as follows. We begin with h and a equal to the "flat" vector v_0 in which every coordinate is equal to 1. We then repeatedly update a_j to be the sum of h_i over all i that point to j, and we update h_j to be the sum of a_i over all i that j points to — this is simply a numerical rendition of our mutually reinforcing relationship among hubs and authorities. In matrix notation, we could write this as follows:

$$a \leftarrow A^T h; \qquad h \leftarrow Aa$$

If we unwind these recurrences, and keep in mind that we normalize both h and a in each iteration to remain a unit vector, then we obtain the following:

$$a = \lim_{n\to\infty} \frac{(A^T A)^n v_0}{|(A^T A)^n v_0|}; \qquad h = \lim_{n\to\infty} \frac{(AA^T)^n v_0}{|(AA^T)^n v_0|}$$

Finally, we can determine these limits explicitly via the following standard theorem about eigenvectors.

Theorem 4 (See Golub and Van Loan [38]) *Let $M \neq 0$ be a symmetric matrix, let λ^* denote the eigenvalue of M with maximum absolute value, and let u_0 denote a vector that is not orthogonal to the subspace consisting of the eigenvectors associated with λ^*. Then the unit vector in the direction of $M^n u_0$ converges to an eigenvector associated with with λ^* as n increases without bound.*

Thus, the limiting vectors of hub and authority weights are eigenvectors of AA^T and $A^T A$ respectively, associated with eigenvalues of maximum absolute value.

In a number of different implementations and studies, both the page-rank and hub/authority methodologies have been shown to provide qualitatively good search results for broad query topics on the WWW — such topics can implicitly involve an underlying set of several million relevant pages. Because the methods are heavily based on link information, they offer another means of circumventing problems that can arise from too great a reliance on pure term-matching techniques.

Although both can be analyzed in the language of eigenvectors, it is interesting to contrast the "one-level" type of influence propagation manifested by Pinski-Narin influence weights [58] with the "two-level" conferral of authority that forms the basis of Kleinberg's hub/authority model [43]. One could argue that the distinctions between the two directly parallel fundamental differences in the social organizations of the scientific literature and the World Wide Web. 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 — for example, `www.honda.com` and `www.toyota.com` are both authoritative sources for the topic "automobile manufacturers," but they will not link to one another for commercial and competitive reasons. 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; a model of the WWW involving both hubs and authorities takes this into account. Such a 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.

A set of hubs densely linking to a common set of authorities can be viewed as a natural graph-theoretic "community" structure associated with a topic of general interest on the Web. Recent work has advanced the argument that this basic type of linkage pattern is a recurring and fundamental structural feature of the World Wide Web; see [37, 46] for further details.

7 Further directions in link-based analysis

We now consider some recent extensions to these link-based approaches. First, Bharat and Henzinger [7] consider modifying the binary link weights used in computing WWW hubs and authorities to incorporate term weighting schemes, thus including text frequency statistics in the underlying iterative algorithm. They also modify link weights to limit the extent to which authority from pages on a single "site" could be conferred to any individual page.

The CLEVER system [12] also builds on the algorithmic framework of hubs and authorities, and includes a number of extensions based on both content and link information. We give two examples. First, it develops a heuristic approach to prevent "topic drift" on large hub pages with many links; this is in response to the problem that on a page containing a large number of links, it is likely that all the links do not focus on a single topic. In such situations it becomes advantageous to treat contiguous subsets of links as mini-hubs, or *pagelets*; one may develop a hub score for such pagelets, down to the level of single links, that participate in the iterations as full-fledged entities. The thesis is that contiguous sets of links on a hub page are more focused on a single topic than the entire page. For instance, a complete hub page

on automobile racing may be a good hub for the topic of "cars", but a contiguous set of links on the page may cater to "Indy racing."

Another extension proposed in [12] makes use of the text that surrounds hyperlink definitions (href's) in Web pages, often referred to as *anchor text* (e.g. McBryan [53]). The use of anchor text in this setting, to weight the links along which authority is propagated, is based on the following observation. When one seeks authoritative pages on mountain climbing, for instance, one might reasonably expect to find the phrase "mountain climbing" in the vicinity of the *tails* – or anchors – of the links pointing to authoritative pages. Thus a hub page containing links to mountain climbing resources, and links to kayaking resources, is likely to contain the phrase "mountain climbing" around the href's pointing to the correct resources resources, but not near the kayaking links. To this end, one boosts the weights of links which occur near instances of query terms. The details of this are given in [11, 12].

8 Conclusion

We have focused primarily on linear algebraic methods that make use of eigenvectors and the singular value decomposition; and from this perspective, we have seen a variety of ways in which methods from linear algebra can be brought to bear on problems in information retrieval and hypertext analysis. The application of such techniques has been made possible by the long-standing use of vector representations for documents in information retrieval, and the deep connections that exist between the combinatorics of link structures and the eigenvectors of their adjacency matrices. We feel that a recurring theme in these areas is the versatility of spectral methods, and the diversity of ways in which eigenvalues and singular values naturally arise in these domains; this offers every indication that spectral methods will remain a significant source of useful techniques for approaching problems in information retrieval.

References

[1] D. Baker and A. McCallum. Distributional clustering of words for text classification. In *Proceedings of ACM Conf. Res. and Development in Information Retrieval*, pages 96–103, 1998.

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

[3] M. Berry. Large scale singular value computations. *International Journal of Supercomputer Applications*, 6:13–49, 1992.

[4] M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4), 1995, 573-595.

[5] M. Berry, S. Dumais, and A. Shippy. A case study of latent semantic indexing. Technical Report CS-95-271, University of Tennessee, January 1995.

[6] M. Berry and R. Fierro. Low-rank orthogonal decompositions for information retrieval applications. Numerical Linear Algebra with Applications 3:4 (1996), pp. 301-328.

[7] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of ACM Conf. Res. and Development in Information Retrieval*, pages 104–111, 1998.

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

[9] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J. Della Pietra, J.C. Lai. Class-based n-gram models of natural language. *Computational Linguistics* 18(1992), pp. 467–479.

[10] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using smart: Trec 4. In D. Harman and E. Vorhees, editors, *The Fourth Text REtrieval Conference (TREC4)*. NIST Special Publication, 1995.

[11] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th World-Wide Web conference*, Amsterdam, 1998. Elsevier Sciences.

[12] S. Chakrabarti, B. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In *SIGIR workshop on hypertext information retrieval*, 1998.

[13] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[14] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[15] W. Donath and A. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay. Clustering in large graphs and matrices. *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[17] S. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236, 1991.

[18] S. Dumais. LSI meets TREC: A status report. In D. Harman, editor, *The First Text REtrieval Conference (TREC1)*, pages 137–152. NIST Special Publication 500-207, 1993.

[19] S. Dumais. Latent semantic indexing (LSI) and

TREC-2. In D. Harman, editor, *The Second Text REtrieval Conference (TREC2)*, pages 105–116. NIST Special Publication 500-215, 1994.

[20] S. Dumais. Using LSI for information filtering: TREC-3 experiments. In D. Harman, editor, *The Third Text REtrieval Conference (TREC3)*. NIST Special Publication, 1995.

[21] S. Dumais, G. Furnas, T. Landauer, and S. Deerwester. Using latent semantic analysis to improve information retrieval. In *CHI88 Proceedings*, pages 281–285, 1988.

[22] S. Dumais, T. Landauer, and M. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 Workshop on Cross-Linguistic Information Retrieval*, pages 16–238, 1996.

[23] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

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

[25] C. Faloutsos and D. Oard. *A Survey of Information Retrieval and Filtering Methods*. Dept. of Computer Science, University of Maryland, 1995.

[26] W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, New York, 1968.

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

[28] P. Foltz and S. Dumais. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.

[29] W. Frakes and editors R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.

[30] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 355–362, 1988.

[31] 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.

[32] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, and R. Harshman. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of ACM Conf. Res. and Development in Information Retrieval*, pages 36–40, 1988.

[33] G. Furnas, T. Landauer, L. Gomez, and S. Dumais. Statistical semantics: Analysis of the potential performance of key-word information systems. *Bell System Technical Journal*, 62(6):1753–1806, 1983.

[34] G. Furnas, T. Landauer, L. Gomez, and S. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

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

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

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

[38] C. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.

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

[40] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1981.

[41] W. Johnson and J. Lindenstrauss. Extension of lipshitz mapping into hilbert space. *Contemp. Math*, 26:189–206, 1984.

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

[43] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998. Also appears as IBM Research Report RJ 10076, May 1997.

[44] University of Singapore KNOW. Machine learning/feature selection bibliography, 1999. http://www.iscs.nus.edu.sg/∼ngkians1/dash.bib.

[45] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 284–292, San Francisco, CA, 1996. Morgan Kaufmann.

[46] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. Trawling emerging cyber-communities automatically. *Proc. 8th International World Wide Web Conference*, 1999.

[47] T. Landauer and S. Dumais. Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[48] T. Landauer, P. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[49] T. Landauer and M. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, Waterloo, Ontario, 1990.

[50] 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.

[51] M. Lesk. The seven ages of information retrieval. In *Proceedings of the Conference for the 50th anniversary of As We May Think*, 1995. available as http://community.bellcore.com/lesk/ages/ages.html.

[52] H.P. Luhn. A Statistical Approach to the Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1:4, October 1957, 309-317.

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

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

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

[56] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of ACM Symposium on Principles of Database Systems*, 1997.

[57] F. Pereira, N. Tishby, L. Lee. Distributional Clustering of English Words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993.

[58] 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).

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

[60] P. Raghavan. Information retrieval algorithms: A survey. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 1997.

[61] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[62] G. Salton. *Automatic Information Organization and Retrieval*. McGraw Hill, New York, 1968.

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

[64] Salton, G. and Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5), 513–23.

[65] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[66] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, MA, 1989.

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

[68] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, March 1972.

[69] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1993.

[70] Text REtrieval Conference. http://trec.nist.gov/.

[71] P. Turney. Feature selection bibliography, 1999. http://ai.iit.nrc.ca/bibliographies/feature-selection.html.

[72] Y. Yang and J. Pederson. Feature selection in statistical learning of text categorization. In *Proceedings of ICML'97*, pages 412–420, 1997.