Topic distillation and spectral filtering

Soumen Chakrabarti Byron E. Dom[*] David Gibson Ravi Kumar
Prabhakar Raghavan Sridhar Rajagopalan
Andrew Tomkins

IBM Research Division
Almaden Research Center
650 Harry Rd.
San Jose, CA 95120-6099
{soumen,dom,gibson,ravi,pragh,sridhar}@almaden.ibm.com

June 29, 1998

Abstract

This paper discusses *topic distillation*, an information retrieval problem that is emerging as a critical task for the www. Algorithms for this problem must distill a small number of high-quality documents addressing a broad topic from a large set of candidates.

We give a review of the literature, and compare the problem with related tasks such as classification, clustering, and indexing. We then describe a general approach to topic distillation with applications to searching and partitioning, based on the algebraic properties of matrices derived from particular documents within the corpus. Our method — which we call *spectral filtering* — combines the use of terms, hyperlinks and anchor-text to improve retrieval performance. We give results for broad-topic queries on the www, and also give some anecdotal results applying the same techniques to US Supreme Court law cases, US patents, and a set of Wall Street Journal newspaper articles.

[*] author to whom all correspondence should be addressed

1 Introduction

This paper discusses a new information retrieval problem known as *topic distillation*,[1] that is emerging as a critical task for the WWW. Algorithms for this problem must distill a small number of high-quality documents (web pages) that address a particular broad topic from a large set of candidates. We give an overview of the problem, describe existing approaches, give detailed comparisons to traditional problems from IR and other fields, and then present results for *spectral filtering*, our approach to the problem. The remainder of the introduction motivates topic distillation as a distinct information retrieval problem that arises in practice and requires new techniques.

With the rapid growth of the WWW, hundreds of millions of documents of varying quality have been made available online to millions of daily users in decentralized fashion. Initially, reflecting the distributed nature of the web, the only facilities for finding pages were browsing and bookmarking. Soon however, in response to the need for centralized page-location services, a number of independent *search engines* appeared. These engines represent the primary approach to information discovery on today's web; good ones are capable of servicing in excess of twenty million queries per day with sub-second average response times.

However, the scope of the web and the diverse body of users means that the same engines must service queries ranging from "What is pin 4 of a 74LS00 TTL NAND gate?" to "Tell me about War?" Responses to specific queries such as the first one tend to be reasonable, especially for users experienced in query construction, and traditional information retrieval techniques perform well. Unfortunately, the situation for general broad-topic queries is worse. As an example, consider the query "fishing." This query hits in more than a million pages on today's web. Top responses from the two largest search engines at the time of this writing tend to be advertising pages, often for companies located on individual lakes or regions — a number of excellent online fishing resources are not ranked highly. A topic distillation algorithm should be able to filter out the large number of irrelevant pages (i.e., pages that mention "fishing for excuses"), low-usefulness pages (i.e., pages about "fishing in the south of Medford, North Dakota" which would be appropriate responses to a more specific query), and low-quality pages (i.e., the large number of advertising pages offering the same products for the same prices) to return thirty great pages about fishing in general.

Some recent work suggests that topic distillation algorithms can perform substantially better than traditional keyword searches for broad topic queries. A recent study [3] collected responses from 37 users over a range of 27 broad-topic queries about pages returned by a traditional search engine, a manually-created resource site, and a topic distillation algorithm similar to the algorithms in this paper. The study concluded that for broad topic searches, users typically rated pages from the traditional search engine as "fair" to "bad" in quality and relevance, while the same users rated pages from the topic distillation algorithm substantially better — better even than pages from the manually-created resource site.

But do queries requiring distillation actually arise on today's web? We found that of queries submitted to the *metacrawler* search engine (http://www.metacrawler.com), almost half (43%) occur in more than 10,000 pages (since a user cannot comfortably scan 10,000

[1] As far as we know, the authors of [7] first coined the phrase "topic distillation" to refer to a problem studied in [36] and [11].

documents, some form of distillation is necessary); about 17% occur in more than 100,000 pages; and 3-4% occur in more than a million pages.[2] As the web continues to grow, it is clear that many queries will match a large number of pages, and based on the web of today it seems unlikely that all or even most of these pages will be high-quality and relevant to the search topic. This suggests that distillation will continue to arise as an important problem.

Finally, we have defined topic distillation informally as search within a large corpus for a small number of high-quality documents addressing a broad topic. The operational defintion of "high-quality" we use to evaluate algorithms is "rated as high-quality with respect to a given query by an unbiased evaluator." In order to approximate this evaluation function, our algorithms consider the content of a page along with any endorsements of the page provided by incoming hyperlinks. These endorsements are weighted by characteristics of the hyperlink, and estimates of the credibility of the page that created the hyperlink.

We call our approach "spectral filtering" because it is based on the spectral properties of matrices derived from relationships within a corpus. It is (1) general — it can be applied to a diverse set of traditional and hyperlinked corpora; (2) flexible — it performs topic distillation, but also supports a variety of other IR tasks including search and clustering; and (3) fast — the computational bottlenecks associated with numerical and algebraic methods such as LSI can be bypassed in our setting.

Section 2 describes existing approaches to topic distillation, and also contains a discussion of related problems and techniques that are appropriate for these problems. Section 3 describes spectral filtering, and Section 4 gives some experimental results.

2 Related work

In reviewing relevant prior work we will address two general problems: (1) quality ranking of documents similar to that performed in topic distillation; and (2) structure discovery in document collections. We discuss the latter because many of the techniques used there appear to be potentially useful in attacking the topic distillation problem also. These two problems are intermingled in the following discussion which is organized by the type of information used.

2.1 Using only text for quality ranking and structure discovery

The bulk of the work and literature in information retrieval has been about the use of only the document's text. See for example NIST's TREC[56] series of annual conferences and their proceedings, or the excellent texts [43] and [46]. When only the text and no citations or other links are available, relevance is often used as an approximation of quality. Attempts to incorporate other dimension(s) of quality tend to involve either simplistic heuristics such as counting the number of words, or sophisticated natural language understanding.

[2] Metacrawler exposes a random subset of the queries it receives. To determine how many pages contain a given query, we submitted the query to the *hotbot* search engine (www.hotbot.com) since its coverage is currently considered to be the largest (see http://www.searchenginewatch.com). Hotbot's coverage is estimated [12] to be around 40% of the web, suggesting that the numbers above may be a substantial underestimate of the number of web pages matching the query.

Text alone has been used in the structure discovery problem also. This is most commonly carried out as a clustering exercise in term-frequency space. As described in Section 3, our approach makes only limited use of document text.

2.2 The use of bibliographic citations

Since long before the advent of hypertext, there has been an established form of explicit directed inter-document links — bibliographic citations, most common in the scientific literature. The field of study that analyzes these citation patterns is known as "bibliometrics" (See the reviews [61] and [39].). Work in this field is focused on exploiting structure characterized by the following mutually dual similarity measures between two documents: "bibliographic coupling"(the number of common citations they contain[58]) and "co-citation" (the frequency with which they both appear as citations in the same document[53].) Larson[38] performs a straightforward application of bibliometric analysis techniques to a collection of web pages corresponding to a particular query, resulting in the discovery of five clusters of related pages.

In addition to the exclusive use of citations some researchers have used both text and citations in analyzing document collections. Shaw [51, 52], for example, uses text and links to perform a graph-based clustering of a document collection.

While there is a strong similarity between the roles of bibliographic citations and hyperlinks, there are also many differences. For example, many hyperlinks are purely navigational aids and don't confer any endorsement of the type we wish to utilize in computing an estimate of a document's quality. Likewise, while bibliographic analysis techniques bear some resemblance to ours, the complexity of the web requires an associated complexity in analysis beyond that used in bibliometrics.

2.3 The use of hyperlinks

The advent of hypertext changed the nature of document collections. A problem was created for information retrieval and at least a partial means to its solution was provided. The problem is that documents tend to be divided into smaller pieces (pages) leaving fewer words with which to assess relevance. Hyperlinks on the other hand, provide a means of incorporating information from neighboring pages. In certain contexts (most notably the web) they also provide information similar to that inherent in bibliographic citations in scientific literature — a kind of endorsement. In the following discussion we review various approaches to utilizing hyperlinks in information retrieval. In those cases where the approach is more relevant to the problem we consider it will be discussed in more detail.

Croft and Turtle[18] propose a scheme for incorporating hypertext links as well as bibliographic citations into an information retrieval system in which the relevance of a document to a query is computed by a Bayesian inference network. This approach is concerned with relevance and hyperlinks are treated by adding corresponding "evidence" nodes to the network. These nodes allow the terms contained in neighboring documents to influence its assessed relevance.

Savoy[47, 48, 49] describes a family of relevance-ranking schemes for doing query-based information retrieval in hypertext. The scheme uses both a term-based inverted index and links that can be either bibliographic citations, hyperlinks or both. Certain aspects of this approach resemble ours. For example the acquisition and expansion of the root set are virtually identical. For the actual ranking, however, *spreading activation*[17] is used. The starting activation values are computed based on linguistic similarity to the query.

2.3.1 Hyper-information

A scheme similar in spirit to those of Savoy but different in its details is that of Marchiori[40]. It is also different in that it is based philosophically on hypertext and implemented in that domain. Thus all links are treated as hyperlinks with the appropriate navigational character (i.e. “clickable”). Given a document (page) in a hypertext collection and a query a relevance rank (“Hyper-information”) is computed for the document with respect to the query. This measure is designed to capture quantitatively the idea that the relevance of a page to a query topic is determined both by the textual information it contains and by the information of the pages it points to by means of hyperlinks. Here “information” refers to a kind of combined quantity/quality measure. This information measure is then applied to rank search engine results.

Let $T(p)$ denote the information measure of only the text contained in page p. A simple example of $T(p)$ is the number of query terms contained in a page p. In a sense, one gets a different Hyper-information measure for every different form for T. Let $H(p)$ denote the *Hyper-information* quantity of page p. This quantity is equal to the page's text information plus a “fading factor” $(0 < F < 1)$ times the sum of the hyper-information of all the pages it points to. If the set of these neighbors is denoted $N(p)$ then formally:

$$H(p) = T(p) + F \sum_{r \in N(p)} H(r).$$

Note that this recursive definition implies that the text information of a page k links away from p is “faded” by a factor of F^k before it is added to $H(p)$.

There are modifications to the definition as thus far presented. These deal with non-idealities of the web. For example, loops (paths of out-links that lead back to the original page) are broken and multiple links from one page to another are counted as one. Basically the idea is that the text information $T(r)$ in a page r reachable from p is only counted once and it is faded by the number of links in the shortest path from p to r. Of course it is only practical to go to a finite depth (path length) in incorporating other pages. Next, in the case of *frame* links and other presentations that pull multiple pages into a single browser view, the union of all the associated text is used in computing $T(p)$. And finally, two different values of F are used: F_{in} for intra-site links and F_{out} for inter-site links.

2.3.2 Supervised Hypertext Categorization

Chakrabarti, Dom and Indyk [5] building on previous work in document (text) categorization[4], devise a scheme for hypertext categorization. In the learning phase parameters of two Bernoulli/Multinomial models are learned from a collection of sample documents. The first models the probability of observing certain terms in a document given its category and the second models the probability of observing certain neighbor-document (one link away) categories. The latter has different parameters for neighbors attached by in-links than for those attached by out-links. The authors experiment with several variations of their scheme. Two of particular interest are:

- A collection of hypertext pages are categorized concurrently using an iterative relaxation-labeling algorithm wherein category-membership probability estimates are propagated across links so that, at each iteration step the current probability estimate for a page is computed using its neighbors' estimates of their own categories from the previous iteration. The initial probability estimates are computed using only the pages' text.

- *Bridges*: In one variation evidence similar in spirit to bibliometric *co-citation* is used. This is based on the observation that if a page (referred to as an "IO Bridge" in [5]) cites (links to) two pages, the probability that those pages are of the same category is greater than would otherwise be expected. The authors observe that the strength of this relationship is dependent on how close the two HREFs are to each other in the IO bridge (citing document). The dependence of this same-category probability as a function of separation was measured for a collection of web documents and the results were used in computing the associated probabilities.

2.3.3 Complexities of hypertext and the World Wide Web

The techniques described thus far implicitly view all pages as the same in a certain sense. That is, while they may differ in what they are *about* and in their numbers of in-links and out-links, they are all treated as functionally the same. In many contexts this is the correct view. In hypertext, however, and especially on the web, the situation is much more complex. In an attempt to deal with this complexity Pirolli, Pitkow, and Rao[42] address the problem of identifying aggregates of pages that correspond to a conceptually unified entity. They use (among other things) link structure, usage paths (taken from server logs), text and meta information about pages. They extract structure from a "web locality", using the following steps.

1. Classification of web pages into eight types which include such categories as: "head" (normal user entry points into the "web locality" and include primarily things like personal and organizational home pages), "indices" (things like tables of contents) and "content" (pages that contain the actual information that users seek). The classification is performed with a linear classifier using the following meta attributes: (1) page size, (2) local in-links, (3) local out-links, (4) frequency of request (from server logs) and (5) the relationship between a page and its children in terms of both textual similarity and the frequency with which the associated links are traversed.

2. The implicit construction of a network linking all the pages. The links and their associated weights are constructed based on hyperlinks, text similarity and user trajectories (click trails).

3. Performing a "spread of activation" over this network. By giving initial activation only to those pages classified as "source indices" which the authors define as "entry points and indices into a related information space".

4. The final aggregation of web pages is based on the final distribution of activation. They look for different distribution patterns depending on what they seek to achieve in solving a particular problem.

Rivlin *et.al.*[14, 15, 44], address a similar issue in the context of providing the user with better navigational aids for hypertext. They introduce the notion of index (high out-link count) and reference (high in-link count) nodes, similar to the HITS notions of hubs and authorities described below. They also use various graph-distance metrics to identify candidate "root" nodes (entry points for groups of pages) and to cluster hypertext pages.

Weiss *et al*[60] describe *HyPursuit*, a system for browsing and searching hypertext collections. The system, which provides several features, is based on an organization of the collection of pages into clusters. They group the pages into a cluster hierarchy using a greedy, bottom-up merging algorithm that makes merging decisions based on a similarity measure that incorporates both text and links. The text component is the usual inner product (a.k.a. "cosine measure") between term-weight vectors. The link-based similarity component is a weighted sum of three terms: the first is based on the lengths of the shortest paths (in both directions) between the two documents, the second is called "Common Ancestors", which is essentially *co-citation* and and the third is called "Common Descendants", which is essentially *bibliographic coupling.*

2.3.4 PageRank

Page[8] describes a technique (used in [9]) that takes a more complex approach than the simple in-link and out-link counting used in [14, 15, 44]. He describes his ranking algorithm as simulating a kind of random walk over the web taken by a web surfer. Assuming that each node (page) is equally likely as a starting point for the walk, the steady-state probability for the surfer being at any node is calculated and the pages are ranked by these probability values. Because this random walk uses the link structure, it is hoped that these steady-state probabilities capture the endorsement implicit in hyperlinks.

The basic idea is that, given that the web surfer is at a particular node at one step in the walk, the probability of being at one of the nodes pointed to (out-links) by that node is equal to one over the number of out-links from that node, while the probability of being at a node not pointed to by that node is zero at the next step. Let A be the transition-probability matrix for this (1st-order Markov) process. Then $A[u][v] = 0$ if there's no link $u \rightarrow v$ and $A[u][v] = 1/n_u$ if there is a link $u \rightarrow v$, where n_u is the number of out-links from node u. If the probability of the surfer being at node v at time step t is given by $p_t[v]$ (letting p_t denote the entire probability vector) and if p_0 represents the initial probability vector, then we have:

$$p_t \quad = \quad A^t p_0,$$

where A^t represents the matrix product $A \times A \times A \times \ldots \times A$ (where A appears t times) and not A-transpose, which we represent by A^T. From this it is clear that the steady-state probability vector is given by:

$$p_\infty \quad = \quad \lim_{t \to \infty} A^t p_0. \tag{1}$$

This can be solved by starting with $p_0[u] = \frac{1}{n}$ for all u and iterating as follows:

$$p_t \quad = \quad A p_{t-1}, \tag{2}$$

if it converges. If it does converge, it converges to an eigenvector of A and it can be shown that it converges to the *principal* (associated with the largest eigenvalue) eigenvector as long as p_0 is not orthogonal to that eigenvector.

There is a problem with this model, however, which might be described as follows. There is always a possibility that our surfer might decide to jump to some page not pointed to by the node he is currently occupying. This will certainly happen if he gets caught in a "trap" - a subgraph from which there is no escape (at least a theoretical possibility). There may be convergence problems for equation (2) associated with such artifacts.

PageRank addresses this problem by assuming that there is some small probability that the surfer jumps to some random page and assuming that all pages are equally likely as targets of such a random jump. Considered as an isolated process the probabilities for this uniformly-random-jump model are given by:

$$p_t[u] \quad = \quad \frac{1}{n-1}\sum_{v \neq u} p_{t-1}[v] \quad = \quad \frac{1}{n-1}\left(1 - p_{t-1}[u]\right), \tag{3}$$

where n is the total number of nodes (web pages). Because n is so large and the components of p so small ($<< 1$) after a reasonable number of iterations, equation (3) can be approximated by:

$$p_t[u] \quad = \quad \frac{1}{n}.$$

This is combined with the initial transition model by assuming that the probability of taking such a random jump is equal to α. This gives a (scalar) probability updating rule of:

$$p_t[u] \quad = \quad \frac{\alpha}{n} \; + \; (1-\alpha)\sum_{v \to u}\left(\frac{1}{n_v}\right) p_{t-1}[v]$$

Going back to the exact form (equation (3)) of the uniformly-random-jump rule, we can write the Markov transition-probability matrix for the combined process as:

$$B \quad = \quad \frac{\alpha}{n-1}(\mathbf{1} - \mathbf{I}) \; + \; (1-\alpha)A,$$

where $\mathbf{1}$ is an $n \times n$ matrix of all ones and $\mathbf{I}$ is the $n \times n$ identity matrix. Thus the ranking obtained is an ordering of the nodes by the projection of their columns in B onto its principal eigenvector.

This ranking algorithm is deployed in the following scenario. Ideally, the algorithm is run over the entire web. Practically speaking, of course, this means some large fraction of the web. These ranks are then kept as part of a document index. When term-based Boolean queries are run against this index, the results are returned sorted by PageRank.

2.4 The HITS technique and its descendants

2.4.1 HITS

Our technique of *spectral filtering* is a generalization of the HITS[36] scheme devised by Jon Kleinberg. HITS produces two distinct but related types of pages in response to a query

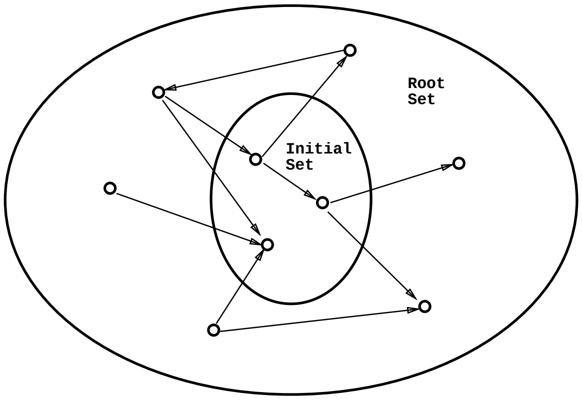

Figure 1: Expanding the initial set into a root set.

topic: *authorities* on the topic (highly-referenced pages), and *hubs* — pages that "point" to many of the authorities. Hubs and authorities exhibit a mutually reinforcing relationship: a good hub points to many good authorities; a good authority is pointed to by many good hubs.[3] HITS proceeds as follows.

1. Starting from a user-supplied query, HITS assembles an *initial set* of pages: typically, up to 200 pages returned by a search engine such as AltaVista [20] on that query. These pages are then expanded to a larger *root set* by adding any pages that are linked to or from any page in the initial set. See Figure 1.

2. HITS associates with each page p a *hub-weight* $h(p)$ and an *authority-weight* $a(p)$, all initialized to 1. Let $p \to q$ denote "page p has a hyperlink to page q". HITS then iteratively updates the h's and a's as follows:

   $$a(p) := \sum_{q \to p} h(q); \qquad h(p) := \sum_{p \to q} a(q).$$

   Thus, a single iteration replaces $a(p)$ by the sum of the $h()$'s of pages pointing to p, and then replaces $h(p)$ by the sum of the $a()$'s of pages pointed to by p.

3. The update operations are performed for all the pages, and the process repeated (normalizing the weights after each iteration).

We now rephrase the iteration in terms of linear algebra. Define matrix $A = [a_{ij}]$ such that a_{ij} is 1 if page i has a link to page j, and 0 otherwise. Then given vectors $\boldsymbol{h}$ and $\boldsymbol{a}$ representing the hub and authority score of each page, the iteration can be re-written as

$$\boldsymbol{h} \leftarrow A\boldsymbol{a}; \qquad \boldsymbol{a} \leftarrow A^T\boldsymbol{h} \tag{4}$$

From classical matrix theory [32], it follows that $\boldsymbol{h}$ converges to the principal eigenvector of AA^T, while $\boldsymbol{a}$ converges to the principal eigenvector of A^TA. Kleinberg further points out that by analogy with *spectral graph partitioning* [21], the *non-principal* eigenvectors of AA^T and A^TA can be used to partition the pages into groups of related hubs and authorities, respectively. (He cites, for instance, the partitioning of pages on "abortion" into pro-choice and pro-life clusters.)

[3] Pages can be both good authorities and good hubs.

2.4.2 HITS refinements

The HITS algorithm as described works well in many cases, but fails in others because it doesn't adequately address the complexities of the web. For example:

- No attempt is made to ensure that the pages acquired in the graph-expansion phase match the query. While this serves a kind of query-expansion function, it often admits unrelated pages, which sometimes results in unrelated pages dominating the highly rated hubs and/or authorities.

- No attempt is made to discriminate between navigational links and those that are the analogs of reference citations. Such links are thus allowed to confer undeserved authority. The minor effect of this is the addition of noise to the the process and results. In the extreme cases most or all of the pages from a single site show up as the top hubs and/or authorities.

- An effect closely related to the navigational-link problem is the existence of certain sites whose associated linkage pattern is pathological with respect to the model assumed by HITS and for which there are an extremely large number of in-links. Most common among these are software companies that sell web browsers (e.g. "This page best viewed with ...") such as Netscape and Microsoft and search engines (and web indices) such as Yahoo![63], AltaVista[20], Excite[23], Infoseek[33] and so on. Such sites frequently get such a high authority rank that they produce a dramatic distortion of what would otherwise be the hub/authority ranking.

- HITS implicitly assumes that all pages are isolated documents. In fact, of course, many hypertext authors divide their hypertext documents into many small pages whereas others write one large page, often with internal labels that can be jumped to via hyperlinks. Unfortunately HTML, at least, provides no mechanism for explicitly declaring a group of pages to be a single hypertext document. There are manifestations of this problem associated with both in-links and out-links.

  - *in-links:* If a single document is divided into many small pages and other referring pages point to individual pages within the document rather than some common entry point like a table of contents, the authority that should go to the document as a whole will be diluted.
  - *out-links:* A large number of pages within a single large document may contain links to the same page outside of the large document. This results in the referred-to page getting a higher authority score than it deserves, which it gives back to the referring pages, thus distorting their hub scores.

ARC: In previous work[11] (ARC) we made a first attempt to address one of HITS's deficiencies - the occasionally exhibited tendency for topics of the pages returned to be different from the query, often a generalization (e.g. a search for the Python scripting language generalized to computer languages - Perl, Java, etc.). In HITS all graph edges (i.e. hyperlinks) receive an equal weight of one. In our modification, however, we weight the edge by a measure of how well the text in the vicinity of the link in the referring page matches the original query because there is a strong tendency for web-page authors to put descriptive text in the vicinity of the link. We count all query terms in a window (whose width w is a parameter) about the link and give the associated edge a weight of one plus this count. The optimum window width was determined by an experiment described in [11]. This weighting scheme can be seen as an attempt to tune the trade-off obtained between recall gained through the query-expansion effect of the graph expansion and precision.

Bharat and Henzinger[7] have concurrently worked on improving the HITS algorithm in several ways. First, they deal with the problem of repeated endorsement from pages on one site (site A) to a single page on different site (site B) (the "out-link" problem) by dividing the associated authority weight for those links by the total number of links from pages on site A to the single page on site B. Second, they also add content analysis of the page text in addition to link analysis: the texts of hubs and authorities that are apparently good are compared with the initial set in a vector-space inner-product sense to obtain a relevance measure. They experiment with using this measure in two ways:

- They prune nodes from the graph before performing the calculations to identify hubs and authorities. They do this based on a computed relevance threshold.
- They multiply the relevance times the page's hub and authority scores during the HITS iteration

and outliers are eliminated. Third, they propose means for controlling the expansion of the initial set using partial content analysis for the purpose of both speed and quality. Heuristics are applied to determine which nodes are likely to have the most influence on the hub/authority calculations. Then a content analysis is performed on only those pages and if found to have insufficient relevance they are eliminated.

3 Spectral filtering

First we develop spectral filtering in Section 3.1 and then discuss some related computational issues in Section 3.2.

3.1 Spectral filtering

We view our method as operating on a domain of *entities*. Initially we consider documents or web pages; later we consider other types of entities. Just as Kleinberg exploits the annotative power latent in hyperlinks, we wish to think more generally in terms of "what does document i say about document j?" To quantify this, we define a (non-negative real-valued) *affinity* a_{ij} from i to j. Then having fixed a set S of entities, we can define the matrix $A = [a_{ij}]$ of directed affinities. At a high level our method consists of three steps:

1. acquisition of the set S of entities to be analyzed. In HITS and in some applications of spectral filtering this process consists of obtaining the root set via a Boolean keyword search and then expanding it to include neighbors (one link distance away);

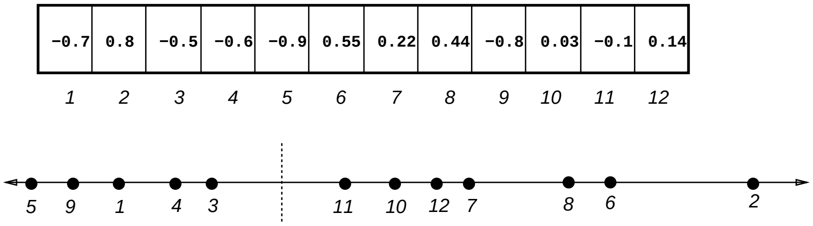

Figure 2: The partitioning process: an eigenvector (top) and the entities ordered on the line (bottom). A split is made between 3 and 11.

2. approximate calculation of one or more of the eigenvectors of one or both of AA^T and A^TA;[4]

3. analysis of the computed eigenvector(s) to rank and/or partition the set of entities.

Note that, in step 2, we perform exactly the iterations of equation 4. We arrive at hub and authority scores converging to the principal eigenvectors (those associated with the largest eigenvalue) of A^TA and AA^T, respectively — we call these *similarity matrices.*[5]

For topic distillation, we perform the above operations on the entities in the subset S. Then, we output the entities with the largest entries in the principal eigenvector of A^TA as the top authorities, and those from the principal eigenvector of AA^T as the top hubs.

But we can also apply spectral filtering for clustering and partitioning either a corpus or a selected subset S. Having set up the matrix A as before, we can also compute the non-principal eigenvectors of A^TA. Because A^TA is real and symmetric, its eigenvectors have real components only. We can view the components of each non-principal eigenvector as assigning to each entity a position on the real line. We deem the entities with large positive values in an eigenvector to be a cluster, and the entities with large negative values to be a different cluster. Alternatively, we can examine the values in the eigenvector (in sorted increasing order). At the largest gap between successive values, we declare a partition into those entities corresponding to values above the gap, and those entities with values below. This is illustrated in Figure 2.

We may view the entries of A^TA as (symmetric) "authority similarities" between entities, and likewise those of AA^T as "hub similarities". Intuitively, the eigenvector operations serve to pull together groups of entities that are all close to one another under the authority (or hub) similarity function.

We now illustrate these ideas with a number of examples. For some examples we give some preliminary and entirely anecdotal evidence to suggest that spectral filtering can be applied in these domains.

[4] Because we compute eigenvectors, the reader may get the impression that the resulting algorithm is slow and impractical for large corpora. In Section 3.2 below we address this concern, showing that in fact we can avoid the exhaustive computation of eigenvectors, for our purposes.

[5] The matrix A as presented contains affinities between entities of the same type. A straightforward generalization gives affinities between entities of different types, e.g., terms and documents. In this case, the rows of A could correspond to terms, and the columns to documents. Although A may not be square, A^TA and AA^T are square and symmetric.

1. HITS: If a_{ij} is 1 if page i links to page j, and 0 otherwise, then our method specializes to HITS.

2. Lexical link weighting: We consider the same topic distillation problem addressed by HITS, but incorporate more information into the affinity matrix. The presence of query terms near a link in document i to document j is suggestive of how the author of document i describes document j. It can thus be used in judging the relevance of document j to a query, as well as in evaluating the connection from i to j with respect to the particular query. In ARC [11] we set a_{ij} proportional to the number of query terms present in the anchor text, and use it to search for web pages relevant to given topics. Section 4 gives extensions to this approach.

3. Term affinities: The preceding examples pertain to hyperlinked corpora; consider now a corpus consisting only of text documents without explicit links. We can define a directed affinity as follows. For documents i, j let $|i \cap j|$ denote the number of terms they have in common. Let $a_{ij} = |i \cap j|/|i|$, where $|i|$ denotes the number of terms in i. Essentially, we are using the commonality of vocabulary to synthesize (weighted) links between documents. As an example, we apply this approach to 47,000 Wall Street Journal articles from 1991. A full comparison of this technique with other relevance ranking algorithms is beyond our scope; we provide instead a few anecdotes to show the types of connections spectral filtering finds:

   - The queries "IBM" and "Microsoft" produce authority lists that coincide in 8 of the top 10 entries. These articles report on the competition and conflict between OS/2 and Windows, then the leading alternatives to MS-DOS. This was one of the most important issues facing either IBM or Microsoft in 1991.
   - The top 10 articles returned by the query "Entertainment" contain two articles about a legal battle between Motown Records and MCA, two articles about leadership change issues in Disney, and three articles about the Nintendo revolution. There were also three general articles describing stock performances in the Entertainment Industry and elsewhere. These were the major Entertainment industry events of the year.
   - The query "Disk Drives" returns a collection of articles describing Sony's emergence as a leader in CD-ROM technology. In addition, there are detailed articles about Maxtor, IBM, and Fujitsu and an article describing how increased storage capacity has contributed to the emergence of multi-media applications.

4. Time-serial corpora: In corpora such as the US Patent database and the Supreme Court rulings, the documents can be thought of as ordered by time (date of creation), and citations only go backwards in time. This is a case where the fact that the iterations go back and forth across (possibly weighted) links is crucial in extracting structure. If, for instance, one were to iterate only along citations (but never in the reverse direction), all the authority would end up in the oldest cases/patents. The field of Bibliometrics (Section 2.2) is also concerned with time-serial corpora. Once again,

rather than provide an extensive comparison, we give a few quick examples of the types of results spectral filtering can provide in these domains.

We consider the database of supreme court rulings, and set the affinity a_{ij} to be 1 if case i cites case j as a precedent, and zero otherwise.[6]

On the search query "right to counsel", the top authorities found were:

Score	Case
0.4425	MIRANDA v. ARIZONA [384 U.S. 436]
0.2464	JOHNSON v. ZERBST [304 U.S. 458]
0.2187	UNITED STATES v. WADE [388 U.S. 218]
0.2174	MASSIAH v. UNITED STATES [377 U.S. 201]
0.1793	POWELL v. STATE OF ALA. [287 U.S. 45]

With the exception of Johnson v. Zerbst, all of the above cases are among the list of landmark cases on Right to Counsel and Self-incrimination in Weinreb [59], a standard legal text on landmark Supreme Court decisions (the fractional numbers are the authority values).

As an example of partitioning, we considered the query "education." The top authorities under the principal eigenvector contain a mixture of cases on various aspects of education such as non-English teaching, freedom of school choice, desegregation and school financing.

When we consider the first non-principal eigenvector, we discover an extremal set of ten cases separated out by this vector. Six of the cases are decisions dealing directly with desegregation of schools. The rest, upon closer examination, turn out to be landmark rulings on the Fourteenth Amendment upon which the case for school desegregation is built; for instance, the first two cases (from 1879) pertain to Fourteenth Amendment rights for colored people as natural-born citizens, and strike down the ability of states to abridge their rights. Thus, one may reasonably infer that the first non-principal "eigenvector" has partitioned out school desegregation cases and their foundations.

We also consider a similar corpus: the US Patent database, available online at [34]. Entities are patents, and affinities are once again citations.

Our first example is the query *cryptography*. The top authorities with scores are given below:

Score	Patent#	Title	Inventors
0.21	4218582	Public key cryptographic...	Hellman, Merkle
0.18	4405829	Cryptographic communications...	Rivest, Shamir, Adleman
0.13	4748668	...identification and signature	Shamir, Fiat

[6] We have also, for instance, increased/reduced the affinities to cases j decided during a selected time-period (e.g., the period during which a particular set of justices is on the bench) to explore the influence of a court's particular leanings — liberal or conservative — on cases decided well after the court's term. However, we do not report these experiments here because of the difficulty of obtaining relevance judgments from the results. This would require legal scholars to study the results, and we have not had the opportunity to do this. We mention the experiments as an example of the flexibility that spectral filtering affords.

We note in passing that the top six authorities sorted by the number of citations to each would be ranked (2,5,1,6,3,4) instead of 1–6, which shows that authority is not equivalent to inlink count. The authorities are celebrated results by academic computer scientists. But what is perhaps more interesting is the nature of the hubs:

Score	Patent#	Title	Inventors
0.55	5455407	Electronic monetary system	Rosen
0.51	5623547	Value transfer system	Jones, Higgins
0.13	5410598	Database...protection system	Shear

These are commercial applications (the assignee of the top hub, for instance, is Citibank) of the basic cryptographic techniques developed by the authorities.

5. Latent Semantic Indexing[22] (LSI) is a dimensionality reduction technique based on SVD[32]. This is performed to capture the "latent semantic structure" of the corpus in question. LSI starts with a term-document matrix, which is a special case of our affinity matrix A (with two different kinds of entities). It then performs a SVD (computing eigenvectors of A^TA and AA^T) and uses the subspace spanned by the first few (say 100) for information retrieval. Both documents and queries are projected into the "document" subspace, where their similarity is measured by (for example) the usual inner product ("cosine measure") between the two projected vectors. The similarity between LSI and spectral filtering is clear; they differ in the way the eigenvectors are used.

6. Collaborative filtering: Consider a setting with two kinds of entities: documents, and people who access them (the precise notion of "access" may be application-dependent: it could mean people who read them with a certain frequency, or people who pay to read them, or people who bookmark them, etc.). For person i and document j, let $a_{ij} = 1$ if i accesses j and 0 otherwise (more generally, a_{ij} could be some non-negative function such as the frequency of access). Now, partitioning using the non-principal eigenvectors would group the people into subsets with similar document-access patterns, and also group the documents into subsets. More generally, the "documents" could be products or other preferences expressed by the people.

Two related themes emerge from these examples, suggesting two broad kinds of applications that can be built from our method. The first uses the authority scores determined by the principal eigenvector to rank entities by their authority on a given topic — this is useful in the context of searching. The second uses non-principal eigenvectors to determine groups of related entities for clustering and hierarchical decomposition. In the remainder of this paper we focus on the former, applying spectral filtering to topic distillation on the WWW.

3.2 Computational issues

The performance of numerical eigenvector computations is often a bottleneck, especially in dealing with large corpora. However, three factors make spectral filtering viable and efficient.

- First, the computation is restricted to a relevant subset of the corpus.

- Second and more important, numerical convergence is not our goal when we wish to rank/group documents by their scores in eigenvectors. Rather, it is the relative ranks of the eigenvector entries that matter. On a search, for instance, we might compute the principal eigenvector and output the ten documents with largest entries in the authority eigenvector (principal eigenvector of $A^T A$). When computing these entries by our iterative methods, the identity of these top ten is usually determined by a very small – often around 5 – iterations, long before numerical convergence is attained. We believe this is an important observation in the use of methods such as ours in information retrieval. Likewise, for the non-principal eigenvectors, we again use iterative methods but stop after 5-10 iterations. Then, the number of operations is typically a small multiple of the number of non-zero entries in the $A^T A$ matrix.

- Finally, the matrix A is typically very sparse, most entries are 0.

4 A small-scale experiment on the WWW

In this section we study broad topic queries on the WWW, and compare spectral filtering against Yahoo! and Infoseek.

This study was performed subsequent to work reported in [11], and differs from the earlier work as follows. First, the earlier study asked users to begin browsing from either a node of a web resource (such as Yahoo!), or from a list of hubs and authorities as provided by a variant of spectral filtering. Users experienced the entire interface provided by the web resource including brief annotative descriptions of each page in a list of links. In the current study, our evaluator began browsing from a list of pages collected from all three sources, with no information about which source contributed which page. We designed the experiment in this way in order to decouple the presentation of a link to a page from that page's inherent quality. Second, the algorithm used in this study benefited from improvements suggested by the results of the earlier work. And finally, estimates of page quality in the current study were provided by a single information specialist rather than a group of arbitrary web users — this decision limited the scope of the study to four queries, but provided higher-quality judgements for each query. We chose the queries carefully to allow all three sources (spectral filtering, Yahoo! and Infoseek) to compete on even footing, as described below.

The goal of the study described here was to show that spectral filtering can provide pages in response to a broad topic query that are comparable in quality to those provided by human experts such as the staff of ontologists at Yahoo!; hence we did not compare to any fully automatic search engines. Having completed this small-scale study described below, we incorporated a number of additional changes suggested by this work, and then began a large-scale evaluation of the resulting algorithm compared to both automatic and human-generated resources. This later study [3] shows that the new algorithm is capable of performing substantially better than automatic search engines, and also better than manually-created resources such as Yahoo!.

4.1 Experiment

Each web page is an entity. Generation of the root set follows the description of Section 2.4. The affinity a_{ij} is 0 if there is no link from page i to page j, and is positive if a link exists. The value of the affinity is a sum of three components. The first component is a default value given to every edge. The second component depends on which of pages i and j fall within the initial set (i.e., which pages contain the query term). The third component has a contribution from each query term. Query terms appearing at distance i within a window of radius W from the hyperlink contribute $W - i$.

We compare spectral filtering to the top two search-engine/indices that have a full taxonomy of categories: Yahoo! and Infoseek.[7] Yahoo! and Infoseek allow users to perform traditional keyword search, or to browse through a hand-created taxonomy of pages. If a user's query happens to have a corresponding node in the taxonomy, the user may then browse from this high-quality starting point, following links that have been classified and inserted by hand. Our goal is to provide automatically-generated sets of links for any query, that correspond to Yahoo! and Infoseek taxonomy nodes. Therefore, we chose query topics for experimentation that correspond to such nodes in Yahoo! and Infoseek. Additionally, we only chose topics whose Yahoo! nodes contained links to what we considered a manageable number ($\sim$ 20) of pages and in which both Yahoo! and Infoseek had, in our opinion, high-quality links.[8]

Spectral filtering ranks pages in two distinct dimensions: as authorities and as hubs. Based on an examination of the number of pages at each node of Yahoo! and Infoseek, we decided to return 20 pages total, and based on the intuition that authorities are more likely to be the eventual goal of a search, we chose to return our top 15 authorities and our top 5 hubs (we consider the relative quality of hubs and authorities below). The query topics are: *Lyme disease*, *telecommuting*, *table tennis*, and *hypertension*.

For each query, the evaluator was given an HTML form containing the query and a simple list of URL titles representing the union of the links provided by Yahoo!, Infoseek, and spectral filtering, each of which was a hyperlink to the page in question. The list was sorted alphabetically by page title. To the left of each title were four check boxes labeled "bad", "fair", "good" and "fantastic". The list was pre-processed by removal of all dead links, and merging of all links that pointed to slightly different variants of the same page. The list contained no indication of which search engines provided which links. The evaluator was free to browse the list at leisure, visiting each page as many time as desired, before deciding on a final quality score.

4.2 Analysis

We converted the allowable ratings "bad," "fair," "good," and "fantastic," into numerical values 0, 1, 2, and 3 respectively. Table 1 gives the average rating of pages returned by each

[7] The data on which this assessment is based are presented at the Search Engine Watch web site (http://searchenginewatch.com/). The top three as of November 1997, as measured by number of visitors, are Yahoo!, Excite and Infoseek in that order. Excite, however, doesn't have a full taxonomy.

[8] Yahoo!'s pages are ordered alphabetically, so rather than imposing an arbitrary ordering and cutoff, we instituted the requirement that the number of links be $\sim$ 20. Infoseek provides an ordered list of links, so we were able to choose the twenty or so top ones.

Search Engine	Average Rating	Data Points
Yahoo	1.50	70
Infoseek	1.73	48
SF	1.52	66
SF Hubs	1.82	17
SF Authorities	1.41	49

Table 1: Average Quality Ratings of Pages, by Search Engine. Quality ratings range from 0 ("bad") to 3 ("fantastic").

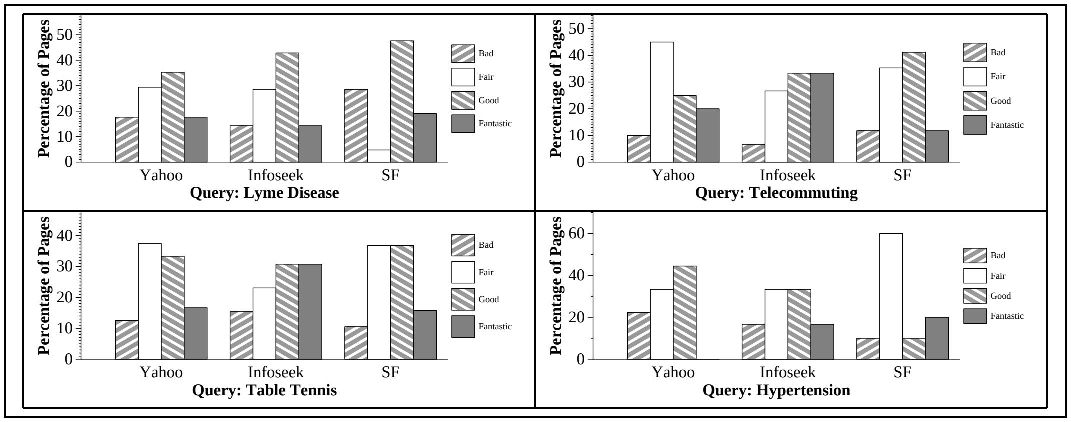

Figure 3: Page quality by search engine

search engine. As the table shows, the differences are not substantial: spectral filtering and Yahoo! return pages ranked respectively 1.52 and 1.50 on average; Infoseek pages are ranked 1.73 on average. However, part of the reason for this difference may be that Infoseek pages contained more dead links than the other two sources, and so only a total of 48 pages were evaluated, as compared to 70 and 66 pages for Yahoo! and SF. Spectral filtering hubs tend to be more highly ranked than any other group, but this is partly because only the top 5 hubs were chosen, rather than the top 15 authorities, and of course average quality tends to decline as the list grows longer (this point is made in more detail below).

4.2.1 Page quality by query

Figure 3 gives histograms showing the number of pages each search engine returned in each category. Across all queries, the fraction of bad pages returned is within $14 \pm 2\%$ for all three systems. Thus, for our query set, spectral filtering is able to remove poor-quality pages as effectively as a human filter. The fraction of pages rated as "good" or "fantastic" was 48% for Yahoo!, 60% for Infoseek, and 54% for SF, indicating that SF is automatically finding high-quality pages as well as the hand-tailored approaches, to within 6% .

4.2.2 Page overlap

Given that the search engines perform similarly, it is natural to ask whether they are finding the same set of pages, or whether each engine is finding a separate set of pages with similar

Rating	Title	Y	I	SF
Fantastic	European Union Co...	✓		
Fantastic	Lyme Disease		✓	
Fantastic	Lyme Disease - Ho...		✓	✓
Fantastic	Lyme Disease Reso...	✓		✓
Fantastic	Lyme Disease Surv...	✓		✓
Fantastic	YPWCnet - Lyme Di...			✓
Good	ACP Online - The ...			✓
Good	American Lyme Dis...	✓	✓	✓
Good	C-T: Net Nature: ...			✓
Good	Health Care Infor...		✓	

Lyme Disease

Rating	Title	Y	I	SF
Fantastic	1995 Telecommutin...		✓	
Fantastic	Escape Artist Tel...	✓		
Fantastic	European Telework...		✓	
Fantastic	Fleming LTD	✓	✓	✓
Fantastic	Smart Valley, Inc...		✓	
Fantastic	Telecommute Ameri...	✓		
Fantastic	Telecommuting, Te...	✓	✓	✓
Good	ATT and Telecommu...			✓
Good	French Telework a...	✓		
Good	General Informati...			✓

Telecommuting

Rating	Title	Y	I	SF
Fantastic	"Table Tennis Onl...		✓	
Fantastic	Bernard Schembri'...	✓		
Fantastic	Gilbert Table Ten...	✓		
Fantastic	International Tab...	✓		✓
Fantastic	Rensselaer Table ...		✓	
Fantastic	Table Tennis Links		✓	
Fantastic	USA Table Tennis	✓		✓
Fantastic	World Wide Ping-Pong		✓	
Fantastic	tt links engels			✓
Good	Bartlesville Tabl...	✓		

Table Tennis

Rating	Title	Y	I	SF
Fantastic	Hypertension Netw...			✓
Fantastic	Hypertension Netw...		✓	✓
Good	American Society ...		✓	
Good	Blood Pressure		✓	
Good	Hypertension, Dia...	✓		✓
Good	Inter-American So...	✓		
Good	Pulmonary Hyperte...	✓		
Good	World Hypertensio...	✓		
Fair	Dr. Palmer's DORO...			✓
Fair	Health Resource D...			✓

Hypertension

Table 2: Search Engine Comparisons on Top-Ranked Pages: Yahoo!, Infoseek, Spectral Filtering

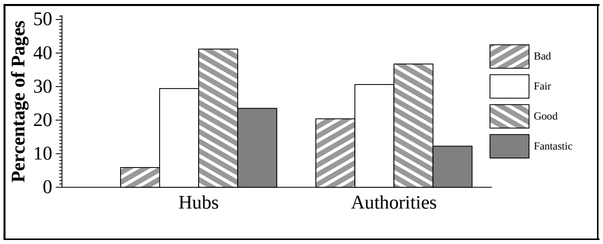

Figure 4: Hubs versus Authorities

quality. For each query, Table 2 lists 10 of the top-ranked pages and shows which search engines found each of these high-quality pages.

4.2.3 Hubs versus authorities

The development of HITS followed from the observation that for broad web queries, users are often interested in the kind of page that has many links (a hub), and the kind of page pointed-to by good hubs (an authority). Pages in the spectral filtering root set are ranked separately as hubs and as authorities. It is natural to ask how actual top-ranked authorities compare to top-ranked hubs in page quality, as determined by our outside expert. Figure 4 presents a histogram comparing hubs and authorities across all queries, using the same format used above to compare different search engines. As the figure shows, the hubs tend to be slightly higher quality, but this difference may be partly due to our choice of 5 hubs versus 15 authorities[9].

[9] Subsequent work [3] actually showed that, in general, hubs retrieved by spectral filtering are more "valuable" than the correponding authorities.

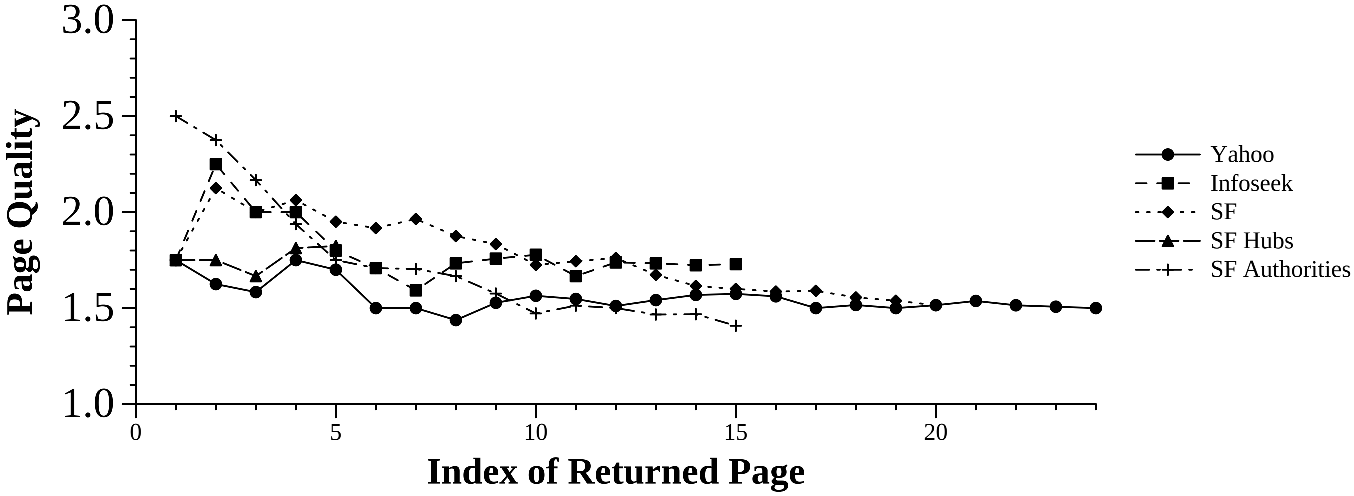

Figure 5: Page Quality as a function of Index. The points at x-value 5, for instance, show the average rating of the first 5 pages returned by each search engine.

4.2.4 Performance by index

The analysis presented so far depends only on the set of pages returned by a particular search engine, not on the ordering of that set. But a search engine with a particular average page quality is much more useful if the top few pages returned are all good, than if early and late pages are of equal quality. Figure 5 shows the average quality of pages returned by a particular search engine from the beginning of returned sequence of pages up to a cutoff point. Results for spectral filtering have been broken into hubs and authorities, but the same data are also presented as an aggregate by interleaving the two sets (choosing the top hub, then the top authority, then the second-best hub, and so on).

5 Conclusions and further work

We describe the emergent *topic distillation* problem for WWW documents. We compare and contrast this problem with related information retrieval problems, and argue that different techniques are effective, motivating the study of topic distillation as a standalone problem. We then present *spectral filtering*, our approach to the problem, and describe a small-scale study of results for pages on the WWW.

Acknowledgment

We would like to thank Lois Nakamura for evaluating the web pages retrieved in the experiments described in Section 4 and to Yael Ravin for her assistance in acquiring the data used in the experiments described in item 3 in Section 3.1.

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] M. Q Wang Baldonado, T. Winograd, "SenseMaker: An information-exploration interface supporting the contextual evaluation of a user's interests," *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1997.

[3] S. Chakrabarti, B.E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Spectral filtering for resource discovery," manuscript, 1998.

[4] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, "Using taxonomy, discriminants, and signatures to navigate in text databases", *23rd International Conference on Very Large Data Bases (VLDB)*. Athens, Greece. 1997.

[5] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext classification using hyperlinks", *ACM SIGMOD Conference on Management of Data*, Seattle, WA, 1998.

[6] 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.

[7] K. Bharat and M.R. Henzinger, "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 469-477. Compressed postscript version: http://www.research.digital.com/SRC/personal/monika/papers/sigir98.ps.gz

[8] Larry Page, PageRank: Bringing Order to the Web. *Stanford Digital Libraries working paper 1997-0072*. 1997. http://www-pcd.stanford.edu/ page/papers/pagerank/index.htm

[9] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th World-Wide Web Conference (WWW7)*, 1998.

[10] B. Bollobás, *Random Graphs*, Academic Press, 1985.

[11] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan and S. Rajagopalan. "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text", *Proceedings of the 7th World-wide web conference (WWW7)*, 1998.

[12] K. Bharat and Andrei Broder. "A technique for measuring the relative size and overlap of public web search engines", *Proceedings of the 7th World-wide web conference (WWW7)*, 1998.

[13] C. Chen. Structuring and visualizing the WWW by generalized similarity analysis. *Proc. 8th ACM Conference on Hypertext*, 177–186, 1997.

[14] Rodrigo A. Botafogo and Ben Shneiderman, "Identifying aggregates in hypertext structures", *Proceedings of ACM Hypertext '91*, pp. 63-74, 1991

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

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

[17] P.R. Cohen and R. Kjeldsen, "Information retrieval by constrained spreading activation in semantic networks", *Information Processing and Management*,**23**, pp. 255-268, 1987

[18] W. Bruce Croft and Howard Turtle, "A retrieval model for incorporating hypertext links", *Proceedings of ACM Hypertext '89*, pp. 213-224, 1989

[19] D. R. Cutting, J. O. Pedersen, D. R. Karger and J. W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Proceedings of ACM SIGIR, 318-329, 1992.

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

[21] 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).

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

[23] Excite Inc., *Excite*, `www.excite.com`.

[24] FindLaw, *FindLaw – LawCrawler*, `www.lawcrawler.com`.

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

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

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

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

[29] K. Fukunaga, *An Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990

[30] R. Furuta, F. M. Shipman III, C. C. Marshall, D. Brenner and H-W. Hsieh. Hypertext paths and the world-wide web: experiences with Walden's paths. *Proc. 8th ACM Conference on Hypertext*, 167–176, 1997.

[31] G. Golovchinsky. What the query told the link: the integration of Hypertext and Information Retrieval. *Proc. 8th ACM Conference on Hypertext*, 67–74, 1997.

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

[33] Infoseek Corporation, *Infoseek search engine*, `www.infoseek.com`.

[34] International Business Machines, *IBM patent server*, `patent.womplex.ibm.com`.

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

[36] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998. Also appears as IBM Research Report RJ 10076(91892) May 1997, and as
`www.cs.cornell.edu/home/kleinber/auth.ps`.

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

[38] 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.

[39] Mengxiong Liu, "Progress in documentation the complexities of citation practice: a review of citation studies", *J. Documentation*, **49**(4), pp.370-408, 1993

[40] Massimo Marchiori, "The Quest for Correct Information on the Web: Hyper Search Engines", *The 6th International World Wide Web Conference (WWW6)*, 1997. Also available at http://atlanta.cs.nchu.edu.tw/www/PAPER222.html.

[41] S. Mukherjea and Y. Hara. Focus+Context Views of World-Wide Web Nodes. *Proc. 8th ACM Conference on Hypertext*, 187–196, 1997.

[42] P. Pirolli, J. Pitkow, R. Rao, "Silk from a sow's ear: Extracting usable structures from the Web," *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1996. (http://www.acm.org:82/sigs/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html)

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

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

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

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

[47] Jaques Savoy, "Searching information in hypertext systems using multiple sources of evidence", *Int. J. Man-Machine Studies*, **38**, pp. 1017-1030, 1993

[48] Jaques Savoy, "An extended vector-processing scheme for searching information in hypertext systems", *Information Processing and Management*,**32**(2), pp. 155-170, 1996.

[49] Jaques Savoy, "Ranking schemes in hybrid boolean systems: a new approach", *J. Am. Soc. Information Sci.*, **48**(3), pp.235-253, 1997

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] TREC - Text REtrieval Conference, co-sponsored by the National Institute of Standards & Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program. (http://trec.nist.gov/)

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

[58] Bella Hass Weinberg, "Bibliographic Coupling: A Review", *Information Storage and Retrieval*, Vol.10, pp. 189-196, 1974

[59] Lloyd L. Weinreb. Leading constitutional cases on criminal justice. Foundation Press, 1982.

[60] 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.

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

[62] Peter Willet, "Recent trends in hierarchical document clustering: a critical review", *Information Processing and Management*, Vol.24, No.5, pp. 577-597, 1988

[63] Yahoo! Corp. *Yahoo!*, `www.yahoo.com`.