Web Search Via Hub Synthesis

Dimitris Achlioptas* Amos Fiat† Anna R. Karlin‡ Frank McSherry§

Small have continual plodders ever won save base authority from others books.
— William Shakespeare, *Loves Labours Lost. Act i. Sc. 1.*
They define themselves in terms of what they oppose.
— George Will, *On conservatives, Newsweek 30 Sep 74.*

Abstract

We present a model for web search that captures in a unified manner three critical components of the problem: how the link structure of the web is generated, how the content of a web document is generated, and how a human searcher generates a query. The key to this unification lies in capturing the correlations between these components in terms of proximity in a shared latent semantic space. Given such a combined model, the correct answer to a search query is well defined, and thus it becomes possible to evaluate web search algorithms rigorously. We present a new web search algorithm, based on spectral techniques, and prove that it is guaranteed to produce an approximately correct answer in our model. The algorithm assumes no knowledge of the model, and is well-defined regardless of the model's accuracy.

1. Introduction

Kleinberg's seminal paper [20] on hubs and authorities introduced a natural paradigm for classifying and ranking web pages, setting off an avalanche of subsequent work [7, 8, 10, 15, 22, 9, 3, 12, 19, 2, 5, 26, 1, 13]. Kleinberg's ideas were implemented in HITS as part of the CLEVER project [7, 10]. Around the same time, Brin and Page [6, 24, 18] developed a highly successful search engine, Google [17], which orders search results according to PageRank, a measure of authority of the underlying page. Both approaches use, chiefly, text matching to determine the set of candidate answers to a query, and then rely on linkage information to define a ranking of these candidate web pages. Unfortunately, as every occasionally frustrated web-searcher can attest, text matching can run into trouble with two ubiquitous phenomena: synonymy (where two distinct terms, like "terrorist" and "freedom-fighter", refer to the same thing) and polysemy (where a single word or phrase, like "bug", has multiple distinct meanings). Synonymy may cause important pages to be overlooked, while polysemy may cause authoritative pages on topics other than the search topic to be included.

Synonymy and polysemy, as problems in information retrieval, long precede the advent of the web and have received a large amount of attention. Latent Semantic Analysis (LSA), pioneered by Deerwester et al. [14], is an idea which has had great practical success addressing these problems. At a high level, LSA embeds a corpus of documents into a low-dimensional "semantic" space by computing a low-rank approximation to the term-document matrix. Computing document/keyword relationships in this low-dimensional space, rather than the original term-document space, appears to address both synonymy and polysemy. By forcing a low-dimensional representation, only those usage patterns that correspond to strong (linear) trends are maintained. Thus, synonymous terms come close in semantic space, as they correlate with other terms in a similar manner, while different meanings of the same word are captured as separate trends since they have nearly orthogonal correlation patterns. Perhaps most importantly, the fact that LSA helps with information retrieval suggests that much of term-usage information can be captured by a simple (low-dimensional) linear model in which each document is expressed as a linear combination of concepts.

In this paper we borrow heavily from earlier work, both on the use of linkage information for determining a document's authority, as well as the use of low-dimensional text representation for addressing synonymy and polysemy. Our main motivation is to give a rigorous framework for web search. To that end, we put forward a generative model which captures in a unified manner three critical components of the problem:

*Microsoft Research, One Microsoft Way, Redmond, WA 98052. optas@microsoft.com

†Dept. of Computer Science, Tel Aviv University, Tel Aviv, Israel. fiat@math.tau.ac.il Work done while on Sabbatical at the University of Washington.

‡Dept. of Computer Science, University of Washington, Seattle, WA 98105. karlin@cs.washington.edu

§Dept. of Computer Science, University of Washington, Seattle, WA 98105. mcsherry@cs.washington.edu

1. How the link structure of the web is generated.
2. How the content of a web document is generated.
3. How a human searcher generates a query.

As we will see, the key to this unification lies in capturing the correlations between these components in terms of proximity in a shared latent semantic space. Given such a combined model, the notion of "the correct answer" to a search query is well defined, and thus it becomes possible to evaluate web search algorithms rigorously.

We present a new web search algorithm we call *SP*[1] and prove rigorous claims about its performance. Specifically, we prove that the algorithm is guaranteed to produce near-optimal results in this model. The algorithm assumes no knowledge of the model, and is well-defined regardless of the model's accuracy.

1.1. Overview and Discussion

In this paper, we define a mathematical model for evaluating web search algorithms. We also present a new web search algorithm, based on spectral techniques, which is guaranteed to give near-optimal results in this model (Theorem 3.1). Our algorithm is entirely motivated by the model and, indeed, our algorithm may not seem intuitive *unless* one considers the model and its implications.

We feel that casting search as a mathematical problem yields a beneficial separation of two intertwined concepts: (a) an abstraction (model) capturing the correlations that make search possible, (b) an algorithm exploiting those correlations. At the very least, such a separation is beneficial in the following sense: if an algorithm is proven to be "good" with respect to a model, yet "no good" in practice, then we will be motivated to further understand in what way the model (and hence the algorithm) is lacking.

The basic idea of our model, which is described in detail in Section 2, is that there exist some number of basic, latent concepts underlying the web (and, thus, most of reality), and that every topic can be represented as a linear combination of these concepts. In turn, web pages, terms, and queries are associated with one or more such topics.

We like to think of the task at hand for *SP* as being:

1. Take the human generated query and determine the topic to which the query refers. There is an infinite set of topics on which humans may generate queries.
2. Synthesize a perfect hub for this topic. The perfect hub is an imaginary page, as no page resembling this imaginary hub need exist. This imaginary hub lists pages in order of decreasing authority on the topic of the query.

[1] Acronym for *SmartyPants*

This task breakdown is explicit in HITS and seems desirable for any search algorithm. Unlike other algorithms, though, in *SP* link information is used both for the second *and* the first subtask. In particular, to find the set of important, relevant documents for the query, we combine the latent semantic analysis of term content with an analysis of link structure. This makes sense in the context of the underlying model in which there is a unified semantic space of link structure, term content and query generation. Thus, we obtain a principled mechanism for avoiding some of the difficulties other search algorithms experience in narrowing down the search space.

The model we assume is rather powerful and allows great flexibility. It may be that some of this flexibility is unnecessary. Certainly, removing this flexibility and simplifying the model does not hurt our results. Thus, if one finds the model too general — say, because the true web does not reflect the full generality of our model — then one can always consider a simplification[2]. Below are some examples of what our model can accommodate:

1. We allow pages to be hubs on one topic and authorities on another. For example, Ron Rivest's home page is a hub for Cryptography but an authority on many subjects (including Ron Rivest and Cryptography).
2. We allow very general term distributions for any topic.
3. We allow distinct term distributions for authorities and hubs on the same topic. For example, hubs on Microsoft may refer to "The Evil Empire" whereas few of Microsoft's own sites use this term.
4. As a page may be a hub on one topic (LINUX) and an authority on another (Microsoft bashing), the terminology used in the page could be a mixture of the hub terminology of LINUX and the authority terminology of Microsoft bashing. This creates great technical difficulties because our goal is to isolate the two components so as to be able to find (say) the best LINUX authorities.

1.2. Additional Related Work

There have been previous attempts to fuse term content and link structure (e.g., [12, 8, 19]). These papers provide experimental evidence that such ideas are valuable and can be implemented. Previous literature makes use of either spectral techniques or the EM algorithm of Dempster. However, whether these algorithms perform well or why remains

[2] In some cases, assuming a simplification may lead to a simpler algorithm.

unclear. The main distinguishing characteristic of our algorithm is that it is provably correct given the model.

Our model is inspired by many previous works, including the term-document model used to rigorously analyze LSI in [25, 2], PLSI [19], the web-link generation model used to study Kleinberg's algorithm [2], PHITS [11] and the combined link structure and term content model of [12]. All of these models, as well as some of the models described in [5] can be viewed as specializations of our model.

1.3. Organization of the Paper

The rest of the paper is organized as follows. Section 2 presents our model. Section 3 presents the new search algorithm. Section 3.4 and the appendix present the proof that the algorithm produces the correct answer. Section 3.4 also discusses extensions and variants to the main theorem. Section 4 concludes with a discussion of the limitations of the model and a number of extensions.

2. The Model

The fundamental assumption in our model is that there exists a set of k unknown (latent) *basic concepts* whose combinations capture every topic considered in the web. How large k is, and what each of these concepts means is unknown. In fact, our algorithm will not (and can not) identify the underlying concepts.

Given such a set of k concepts, a *topic* is a k-dimensional vector w, describing the contribution of each of the basic concepts to this topic; the ratio between the i-th and j-th coordinates of w reflects the relative contributions of the underlying i-th and j-th concepts to this subject. In order to ascribe a probabilistic interpretation to the various quantities in the model, we assume that the coordinates of each topic vector are non-negative.

Associated with each web page p are two vectors:

- The first vector associated with p is a k-tuple $A^{(p)}$ – reflecting the topic on which p is an *authority*. This topic captures the content on which this page is an authority and, therefore, influences the incoming links to this page. Two web pages p and q are authorities on the same topic if $A^{(p)}$ is a scalar multiple of $A^{(q)}$. It is the magnitude of the web page's authority vector, $|A^{(p)}|_1$, which determines how *strong* an authority it is on that particular topic.

- The second vector associated with p is a k-tuple $H^{(p)}$ – reflecting the topic on which p is a *hub*, i.e., the topic that defines the set of links from p to other pages. Intuitively, the hub topic of a page is (usually) pretty well described by the anchor text for the links from that page. As before, two web pages p and q are hubs on the same topic if $H^{(p)}$ is a scalar multiple of $H^{(q)}$.

Remarks:

1. A page's topics represent the union (sum) of all "real-life" topics on the page. Hence, if, for example, page p_c is the concatenation of pages p_1, p_2 then each of the two vectors for p_c is the sum of the corresponding vectors for p_1, p_2.

2. The authority topic and the hub topic of a page may be identical, orthogonal or anything in between.

2.1. Link Generation

Given two pages p and q, our model assumes that the number of links from p to q is a random variable X_{pq} with expected value equal to $\langle H^{(p)}, A^{(q)}\rangle$, the inner product of $H^{(p)}$ with $A^{(q)}$. The intuition is that the more closely aligned the hub topic of page p is with the authority topic of page q, the more likely it is that there will be a link from p to q. In addition, the stronger a hub p is (as measured by the magnitude of $|H^{(p)}|_1$), and/or the stronger an authority q is (as measured by the magnitude of $|A^{(q)}|_1$), the more likely it is that there will be a link from p to q. Our model allows the distribution of X_{pq} to be arbitrary as long as $\mathbf{E}(X_{pq}) = \langle H^{(p)}, A^{(p)}\rangle$, and the range of X_{pq} is bounded by a constant[3] independent of the number of web documents.

Thus, we can describe the link generation model in terms of an n by n matrix W, where n is the number of documents on the web. W is the product of two matrices

$$W = HA^T,$$

where H and A are both $n \times k$ matrices whose rows are indexed by pages. The p-th row of H is $(H^{(p)})^T$ and the p-th row of A is $(A^{(p)})^T$. Each entry in the matrix W represents the expected number of links between the corresponding web documents.

We denote the actual link structure of the web by $\widehat{W}$, where $\widehat{W}[i, j]$ is the number of links between page i and page j. $\widehat{W}$ is an *instantiation* of the random web model defined by the matrix $W = HA^T$. As discussed above, the $[i, j]$ entry of $\widehat{W}$ is obtained by sampling from a distribution with expectation $W[i, j]$ that is of bounded range.

2.2. Term Content

We now introduce the term distributions in web pages. Associated with each term are two distributions:

[3] In practice, the number of links from one page to another rarely exceeds one.

- The first distribution describes the use of the term u as an authoritative term, and is given by a k-tuple $S_A^{(u)}$. The ith entry of this tuple is the expected number of occurrences of the term u in a *pure* authoritative document on concept i that is a hub on nothing, i.e., a web page with authority topic e_i (1 on the ith coordinate, and 0 everywhere else) and hub topic that is 0 everywhere.

- The second distribution describes the use of the term as a hub term (e.g., anchor text) and is given by a k-tuple $S_H^{(u)}$. The ith entry of this tuple is the expected number of occurrences of the term u in a *pure* hub document on concept i that is an authority on nothing.

These distributions can be combined to form two matrices, S_H, the $\ell \times k$ matrix whose rows are indexed by terms, where row u is the vector $(S_H^{(u)})^T$, and S_A, the $\ell \times k$ matrix, whose rows are indexed by terms, where row u is the vector $(S_A^{(u)})^T$.

Our model assumes that terms on a page p with authority topic $A^{(p)}$ and hub topic $H^{(p)}$ are generated from a distribution of bounded range where the expected number of occurrences of term u is

$$\langle A^{(p)}, S_A^{(u)} \rangle + \langle H^{(p)}, S_H^{(u)} \rangle \ .$$

Thus, we can describe the term generation model in terms of an n by ℓ matrix S, where n is the number of documents on the web and ℓ is the total number of terms,

$$S = HS_H^T + AS_A^T \ .$$

The $[i, j]$ entry in S represents the expected number of occurrences of term j in document i.

Analagously to W and $\widehat{W}$, we denote the actual document-term matrix for the web by $\widehat{S}$. The $[i, j]$ entry of $\widehat{S}$ is the number of occurrences of term j in page i, and is assumed by the model to be an instantiation of the the matrix S; the $[i, j]$ entry of $\widehat{S}$ is obtained by sampling from a distribution with expectation $S[i, j]$ of bounded range.

2.3. Human Generated Queries

We assume that the human searcher has in mind some topic on which he wishes to find the most authoritative pages. Intuitively, the terms that the searcher presents to the search engine should be the terms that a perfect hub on this topic would use. A useful think about this is to imagine that the searcher presents the search engine with a portion of the anchor text that a perfect hub would use to describe the links to authorititative pages on this topic.

This motivates our model for the query generation process. In order to generate the search terms of a query:

- The searcher chooses the k-tuple v describing the topic he wishes to search for in terms of the underlying k concepts.

- The searcher computes the vector $q = v^T S_H^T$. Observe that $q[u]$, the u-th entry of q, is the expected number of occurrences of the term u in a pure hub page on topic v.

- The searcher then decides whether or not to include term u among his search terms by sampling from a distribution with expectation $q[u]$. We denote the instantiation of the random process by $\widehat{q}[u]$.

The input to the search engine consists of the terms with nonzero coordinates in the vector $\widehat{q}$.[4]

By choosing the amplitude of v ($|v|_1$) to be very small, the searcher can guarantee that only a small number of search terms will be sampled . In this case, $\widehat{q}$ will be largely zero and have very few non-zero entries.

2.4. The Correct Answer to a Search Query

Given our model as stated above, the searcher is looking for the most authoritative pages on topic v. In our model, the relative authoritativeness of two pages p and q on the topic v is given by the ratio between $\langle v, A^{(p)} \rangle$ and $\langle v, A^{(q)} \rangle$. Thus, *the correct answer to the search query* is given by presenting the user with an authority value for each page, as given by the entries of $v^T A^T$. Ideally, web pages should first be sorted according to their the entries in $v^T A^T$, in descending order, and then presented to the user in this order, the order of decreasing authoritativeness on the topic of interest.

2.5. Our Model versus Google and HITS

To aid our comparison, we begin by arguing that both HITS and Google work correctly if k is 1, i.e., if there is a single concept along which pages are to be ranked and chosen.

Indeed, assume that the first part in the HITS algorithm succeeds in providing a set of candidate documents that correspond to a single topic[5]. Then the link generation model $W = HA^T$ is a rank one matrix ($k = 1$) and, as argued in [2], HITS operates provably correctly and robustly in this setting: HITS computes the right singular vector $\widehat{W}$, where $\widehat{W}$ is the instantiation (real-life version) of W. HITS then

[4]It should be clear that such a subliminal process goes on deep in every user's subconscious...

[5]In HITS there is always a preliminary stage in which documents relevant to the search query are isolated before the algorithm is applied. In our setting, this constitutes an attempt to reduce the document set to a rank 1 set. Thus, here, W refers only to the documents isolated in the first stage of the algorithm.

ranks pages according to their value in this singular vector. The result of the algorithm is, with high probability (over the instantiation process), very close to the ranking associated with the right singular vector of W, which is proportional to the authority vector A.

A similar claim can be made for Google. The PageRank of a page is its stationary probability under a random walk on $\widehat{W}$ which also occasionally jumps to a randomly chosen page. Ignoring this extra randomization (which guarantees that a stationary probability exists and which also speeds up mixing), we observe that the primary left eigenvector of the stochastic matrix associated with W (the Markov chain obtained by dividing every row of W by the sum of the entries in that row) is *equal* to the primary right singular vector of W if W is rank 1.

What this means is that in our simplified versions of Google and HITS algorithms, they both give essentially the *same* correct answer if the rank of the web model is 1.[6] This will in fact be the same result returned by *SP* as well (all terms are equal when the rank of the web is one). In this sense, *SP* generalizes both Google and HITS.

2.6. Summary

We present a unified probabilistic model for link generation and term content in the web and for the process by which a users generate queries. Given the model, the correct answer to the user query is well defined. In Section 3 we will present an algorithm which is guaranteed to produce near-optimal results in this model. We emphasize that the algorithm we describe in Section 3 is well defined regardless of the accuracy of the model, and assumes no knowledge of the model. In fact, we are far from believing that this generative model is in any way related to the unfolding of reality[7]. Rather, we hope that it is good enough an approximation to motivate algorithms that will improve *in practice* over current search engines. The success of [20, 17] gives us hope.

3. The Algorithm *SP*

The algorithm takes as input a search query $\widehat{q}$. Under the model, this query is generated by a human searcher by instantiating $q = v^T S_H^T$ for some topic v. The goal of *SP* is to compute the authoritativeness implied by $v^T A^T$. To do so, *SP* makes use of the web graph $\widehat{W}$ and the web term matrix $\widehat{S}$, both of which are derived by crawling the web. An interesting feature of *SP* is that it does not compute either v or A. In fact, one can show that it is not possible to explicitly derive those matrices given $\widehat{W}$ and $\widehat{S}$ only. Nonetheless, as we will see, *SP* does extract the relative order of the entries of $v^T A^T$.

3.1. Notation and Linear Algebra Facts

For two matrices A and B with an equal number of rows, let $[A|B]$ denote the matrix whose rows are the concatenation of the rows of A and B. We use the notation $\sigma_i(A)$ to denote the i-th largest singular value of a matrix A. We use the notation $[0^n]$ to denote a row vector with n zeros, and the notation $[0^{i\times j}]$ to denote an all zero matrix of dimensions $i \times j$. Finally, we use the singular value decomposition (SVD) of a matrix $B \in R^{n\times m}$, $B = U\Sigma V^T$ where U is a matrix of dimensions $n \times \text{rank}(B)$ whose columns are orthonormal, Σ is a diagonal matrix of dimensions $\text{rank}(B) \times \text{rank}(B)$, and V^T is a matrix of dimensions $\text{rank}(B) \times m$ whose rows are orthonormal. The $[i, i]$ entry of Σ is $\sigma_i(B)$. For an overview of the SVD and its properties, see [16].

We say that $\widehat{B}$ is an *instantiation* of B if each entry $\widehat{B}[i, j]$, independently, is chosen at random from a distribution with a bounded range and expectation $B[i, j]$.

Let $\widehat{W} \in Z^{n\times n}$ be a matrix such that $\widehat{W}[i, j]$ is the number of links between page i to page j. Let $\widehat{S} \in Z^{n\times \ell}$ be the matrix such that $\widehat{S}[i, j]$ is the number of occurrences of term j in page i. Let a query vector $\widehat{q} \in Z^{\ell}$ be the characteristic vector of the query: $\widehat{q}[i]$ is the number of occurrences of term i in the query.

3.2. An Easy Special Case and Intuition

There is an interesting special case of the model[8] where the search problem is particularly easy to solve: when the hub text and the authority text can be easily separated, as they would be, for example, if all of the anchor text on a web page was hub text and all the rest of the text on that page was authority text.

In this scenario, a simple approach is to throw out authority text on each page, and view the document-term matrix as the instantiation of HS_H^T.

To get a feel for why this case is easier (when H, A and S_H are all rank k), imagine that the algorithm has available to it the model matrices $W = HA^T$ and $S = HS_H^T$, and the actual query $q = v^T S_H^T$ instead of their instantiations[9]. The

[6] Of course, in practice they don't give the same answer due to a combination of factors including: (a) Google uses a global rank computation, whereas HITS uses a restricted portion of the web for his computation, (b) Google adds jumps to random locations on the web to the Markov chain, and (c) the web is actually not rank 1.

[7] Almost as far as the reader...and in Section 4.1, we discuss a number of the model's limitations.

[8] Another equally easy special case is when $H = A$, however, as discussed below, there is empirical evidence suggesting that this is unlikely in practice.

[9] Matrix perturbation theory allows us to prove that by using low rank approximations to their instantiations, we can essentially recover these

idea is to find a vector u such that $u^T S^T = q (= v^T S_H^T)$, which is easily done by multiplying the query vector q by the pseudo-inverse of S^T. Given such a vector, we have $u^T S^T = u^T H S_H^T = v^T S_H^T$, which implies that $u^T H = v^T$, since S_H^T is rank k. Finally, by computing $u^T W$, we obtain the result we want, since $u^T W = u^T H A^T = v^T A^T$.

We can now easily give intuition for the algorithm for the general case (again supposing that the algorithm has available to it the model matrices rather than their instantiations). Our goal, once again, is to find a vector u such that $u^T H = v^T$, since from such a vector we can compute the desired result as above. To do this, consider the matrix M obtained by concatenating W^T with S, i.e. $M = (AH^T | HS_H^T + AS_A^T)$. A typical linear combination of rows of this matrix, say $u^T M$, looks like $(y^T H^T | x^T S_H^T + y^T S_A^T)$ for some pair of vectors x and y. What our algorithm does is find a linear combination of rows such that $y^T H^T = 0$ and $x^T S_H^T + y^T S_A^T = q = v^T S_H^T$. Together these imply that $y = 0$ and $x = v$, since the matrices H and S_H are of full rank k. Thus, we have found a u such that $u^T H = v^T$, as needed.

3.3. The Algorithm *SP*

SP performs the following preprocessing (of the entire web) independently of the query:

1. Compute the SVD of the matrix
$$\widehat{M} = [\widehat{W}^T | \widehat{S}] = U_{\widehat{M}} \Sigma_{\widehat{M}} V_{\widehat{M}}^T \ .$$
Note that $\widehat{M} \in R^{n \times (n+\ell)}$ (n is the number of web pages and ℓ is the number of terms).

2. Choose the largest index m such that the difference $|\sigma_m(\widehat{M}) - \sigma_{m+1}(\widehat{M})|$ is sufficiently large (we require $\omega(\sqrt{n+\ell})$). Let $\widehat{M}_m = U_{\widehat{M}_m} \Sigma_{\widehat{M}_m} V_{\widehat{M}_m}^T$ be the rank m SVD approximation to $\widehat{M}$.

3. Compute the SVD of the matrix $\widehat{W} = U_{\widehat{W}} \Sigma_{\widehat{W}} V_{\widehat{W}}^T$.

4. Choose the largest index r such that the difference $|\sigma_r(\widehat{W}) - \sigma_{r+1}(\widehat{W})|$ is sufficiently large ($\omega(\sqrt{n})$). Let $\widehat{W}_r = U_{\widehat{W}_r} \Sigma_{\widehat{W}_r} V_{\widehat{W}_r}$ be the rank r SVD approximation to $\widehat{W}$.

Once a query vector $\widehat{q} \in R^\ell$ is presented, let $\widehat{q}'^T = [0^n | \widehat{q}^T] \in R^{n+\ell}$. *SP* does the following query-dependent computation:

- Compute the vector
$$w = \widehat{q}'^T \widehat{M}_m^{-1} \widehat{W}_r,$$
where $\widehat{M}_m^{-1} = V_{\widehat{M}_m} \Sigma_{\widehat{M}_m}^{-1} U_{\widehat{M}_m}^T$ is the pseudo inverse of $\widehat{M}_m$.

- The search result is a value $w(p)$ for each web page p, where $w(p)$ is the authoritativeness of p on the query topic[10],

3.4. The Main Theorem

For any matrix B, let $r_i(B) \geq 1$ denote the ratio between the primary singular value and the i-th singular value of B: $r_i(B) = \sigma_1(B)/\sigma_i(B)$. If $r(B) = 1$ then this means that the singular values do not drop at all, the larger $r_i(B)$ is the larger the drop in singular values.

Our main goal is now to prove the following theorem:

Theorem 3.1. *Given that the web link structure $\widehat{W}$, term content $\widehat{S}$, and query $\widehat{q}$, are generated according to our model ($\widehat{W}$ is an instantiation of $W = HA^T$, $\widehat{S}$ is an instantiation of $S = AS_A^T + HS_H^T$, and $\widehat{q}$ is an instantiation of $q = v^T S_H^T$), and given that*

- *$\widehat{q}$ has $\omega(k \cdot r_k(W)^2 r_{2k}(M)^2)$ terms,*
- *$\sigma_k(W) \in \omega(r_{2k}(M)\sqrt{n})$ and $\sigma_{2k}(M) \in \omega(r_k(W) r_{2k}(M)\sqrt{n})$,*
- *W and HS_A^T are rank k, $M = [W^T | S]$ is rank $2k$, $\ell = O(n)$,*

then SP computes a vector of authoritativeness that is very close to the correct answer:

$$\frac{\|\widehat{q}'^T \widehat{M}_m^{-1} \widehat{W}_r - v^T A^T\|_2}{\|v^T A^T\|_2} \in o(1). \qquad (1)$$

The condition that M is of rank $2k$ seems to be fairly reasonable. It essentially says that the hub structure and the authority structure of the web are sufficiently different from one another. If indeed, as empirical evidence suggests (e.g., [21]), much of the hub and authority structure of the web is defined by dense, directed bipartite subgraphs, then this assumption is likely to hold. In Section 4.2 we present an algorithm for the case where the rank of M is less than $2k$.

As for the conditions on the singular values of W and M: if, for example, all the singular values of W and M are $\Theta(n)$, then all the conditions of the theorem are met, and the query needs to have only a constant number of terms. At the

model matrices from their instantiations if the singular values of the matrices involved are sufficiently large.

[10] Of course, web pages should be presented to the user in the order of decreasing authoritativeness.

other extreme, if the singular values of W and M drop off so that $\sigma_1(W) \in O(n)$, $\sigma_k(W) \in \omega(n^{5/6})$, $\sigma_1(M) \in O(n)$, $\sigma_{2k}(M) \in \omega(n^{5/6})$ then the conditions of the theorem are also met, but in this case, a ridiculously large number of search terms are required in order to guarantee that the correct answer is found.

The proof of Theorem 3.1 is given in the appendix. For lack of space, many details are omitted.

We can also give a stronger version of this theorem (Theorem 3.2) when we allow the model to be corrupted by an error. This version is useful to justify our approach on the real web where we do not expect the correct model to have rank exactly k, but rather we expect there to be a long tail of additional singular values.

Theorem 3.2. *Under the conditions of Theorem 3.1, assume we were to corrupt the matrix M by adding an error matrix E, $M' = M + E$, where $\|E\|_2 \in o(|\sigma_{2k}(M) - \sqrt{n}|)$ (this also implies an appropriate corruption of W as $M = [W^T|S]$). If we were then to generate $\widehat{M}$ from M', the claims of the theorem would remain unchanged.*

We omit the proof of this theorem in this extended abstract. It follows the same outline as Theorem 3.1, using Lemma 3.0.3 of [2].

Finally, we note that there is experimental evidence that suggests that the singular values for the web graph and the document-term matrix might be Zipf (i.e., σ_i proportional to n/i) [23, 21]. If we assume this in our model, we can show as a corollary to Theorem 3.2 that our algorithm works correctly using only a constant number of search terms.

4. Discussion and Extensions

4.1. Limitations of the Model

Clearly, the effectiveness of our model as a tool for guiding the design of web search algorithms remains open; an empirical evaluation is the obvious next step in this research. The following are all potentially serious limitations of our model:

1. The model assumes that all correlations are linear.
2. The model assumes that entries in the various matrices are instantiated (based on the probabilistic model) independently.
3. By defining the authoritativeness of a page p on topic v to be $v \cdot A^{(p)}$, a strong authoritative site on a related topic (not precisely aligned with v) may be more authoritative on topic v than a weaker authority whose topic is precisely v.
4. It is essentially impossible to define a page that has a particular hub topic, uses precisely the terminology associated with that hub topic, but rather than pointing to the most authoritative sites on the topic, points to the least authoritative sites on the topic.
5. More generally, there is no notion of a bad hub or a bad authority. Every page has a hub topic and an authority topic and what distinguishes two pages on the same topic is the strength of their authoritativeness or hubbiness on that topic.

For an interesting comparative discussion of the limitations of a variety of web search algorithms, along with an experimental evaluation see [5].

4.2. If Hubs and Authorities cannot be linearly separated

We have assumed that $M = [AH^T|AS_A^T + HS_H^T]$ was rank $2k$ and W was rank k. Of course, when running the algorithm we don't know in advance what k is. What we really need is that there be a large gap between the $2k$-th and $(2k+1)$-st singular values of $\widehat{M}$ and simultaneously a large gap between the k-th and $(k+1)$-st singular values of $\widehat{W}$. In other words, we can actually check that the conditions of the theorem are met.

Suppose, however, that in reality M is not rank $2k$, and we compute the singular values of $\widehat{M}$ and $\widehat{W}$ and find that the correct interpretation of these singular values is that M has some rank $k \le i < 2k$. For example, this would happen if $H = A$, *i.e.*, the hub topic and authority topic of each page is identical.

Consider the following modification to the query-dependent portion of *SP*:

1. Find u' so that $u'^T\widehat{M}_m = [y^T|\widehat{q}^T V_{\widehat{S}_m} V_{\widehat{S}_m}^T]$ and $\|y\|_2$ is minimized, where $\widehat{S}_m$ represents the rightmost ℓ columns of $\widehat{M}_m$.
2. Choose $w = u'\widehat{W}_r$ and output the order given by the coordinates of w.

The significance of M being of rank $< 2k$ is that one may not be able to isolate the pure hub properties of the topic. In some cases (for some query vectors $q = vS_H^T$), one can only compute the links of pages that have a mixture of hub terms S_H and some authority terms S_A. *I.e.*, we would get $h^T S_H^T + a^T S_A^T = v^T S_H^T$ and *SP* would output $h^T A$ rather than vA.

Our goal in such a case would be to reduce the (possibly misleading) authority element in the synthesized hub as much as possible. That is exactly what is being done in the variant above.

4.3. A Recursive Approach

One problem with the approach taken by *SP* is that the k fundamental concepts in the overall web will be quite different than the k fundamental concepts when we restrict attention to computer science sites.

What we would like to do is to focus on the relevant subset of documents so that we indeed have sufficient resolution to identify the topic at hand, and the topic does not vanish as one of the insignificant singular values.

A natural recursive approach would be to apply *SP*, sort the sites by both authority value and hub value, take the top 1/3 of the most authoritative sites along with the top 1/3 of the hub sites and recur.

Clearly, this process converges. We suspect (but don't have any evidence) that this may work well in practice. If the search term "singular value decomposition" (say) has *any* significance in "Engineering" (say) then the first round will post Engineering authorities and hubs high up in the list and Theology authorities and hubs low in the list. So the 2nd round will start with Theology and other unrelated topics omitted, and a higher resolution set of concepts for Engineering. In fact, since we don't know the resolution at which the user wants to perform the query, it may be useful to provide answers at various resolutions.

References

[1] Brian Amento, Loren G. Terveen, and Willuam C. Hill. Does "authority" mean quality? predicting expert quality ratings of web documents. In *Research and Development in Information Retrieval*, pages 296–303, 2000.

[2] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *STOC 2001*, 2001.

[3] Krishna Bharat and Monika Rauch Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Research and Development in Information Retrieval*, pages 104–111, 1998.

[4] Ravi B. Boppana. Eigenvalues and graph bisection: An average case analysis. In *Proc. 28th Annual FOCS*, 1987.

[5] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Finding authorities and hubs from link structures on the world wide web.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Hypersearching the web. *Scientific American*, June 1999.

[8] S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Spectral filtering for resource discovery, 1998.

[9] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30(1–7):65–74, 1998.

[10] Soumen Chakrabarti, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, and Jon Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[11] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Inn Proceedings of the 17th International Conference on Machine Learning*, 2000.

[12] David Cohn. The missing link - a probabilistic model of document content and hypertext connectivity.

[13] Jeffrey Dean and Monika Rauch Henzinger. Finding related pages in the world wide web. *WWW8 / Computer Networks*, 31(11-16):1467–1479, 1999.

[14] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[15] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[16] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[17] Google. `http://www.google.com`.

[18] Monika Rauch Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. Measuring index quality using random walks on the web. *WWW8 / Computer Networks*, 31(11-16):1291–1303, 1999.

[19] Thomas Hofmann. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, pages 50–57, 1999.

[20] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[21] S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31(11-16):1481–1493, 1999.

[22] Manjara. `http://cluster.cs.yale.edu`.

[23] Chris Ding Nersc. A dual probabilistic model for latent semantic indexing in information retrieval and filtering.

[24] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, 1998.

[25] Christos Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In Proceedings of ACM Symposium on Principles of Database Systems, 1997.

[26] Jan Ua Ry. The pagerank citation ranking: Bringing order to the web.

[27] G.W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

A. Appendix: Proof of Theorem 3.1

The proof has three parts: first we give two alternate formulations for the term $v^T A^T$ in Equation (1). Next, we give a lower bound on $\|v^T A^T\|_2$, and finally an upper bound on $\|\widehat{q}'^T \widehat{M}_m^{-1} \widehat{W}_r - v^T A^T\|_2 / \|v^T A^T\|_2$.

A.0.1 Reformulating $v^T A^T$

Define the matrix $M = [W^T|S] \in R^{n\times(n+\ell)}$, let $q'^T = [0^n|q^T] \in R^{n+\ell}$.

Claim A.1. *q'^T is in the row space of M.*

Proof. We can rewrite M as follows

$$\begin{aligned} M &= [W^T|S] \\ &= [AH^T|AS_A^T + HS_H^T] \\ &= [A|H]\left[\begin{array}{cc} H^T & S_A^T \\ 0^{k\times n} & S_H^T \end{array}\right]. \end{aligned} \tag{2}$$

If rank$(M) = 2k$ then the row space of M is equal to the row space of the right matrix in equation (2). Recall that $v \in R^k$ (chosen by the searcher) is such that $q^T = v^T S_H^T$. This implies that $q'^T = [0^n|q^T] = [0^n|v^T S_H^T] = v^T[0^{k\times n}|S_H^T]$ is in the row space of M. □

Therefore, there exists some $u \in R^n$ in the column space of M such that $u^T M = q'^T$.

Claim A.2. *Let $u \in R^n$ be such that $u^T M = q'^T$, then $u^T W = v^T A^T$.*

Proof. We can write $M = [AH^T|HS_H^T + AS_A^T]$, we know that $u^T M = q'^T = [0^n|q^T]$. From this we learn that $u^T AH^T = 0$. As H^T is rank k it follows that $u^T A = [0^k]$. Thus, $u^T M = [0^n|u^T HS_H^T] = q'^T$ which implies that $u^T HS_H^T = q = v^T S_H^T$. As rank$(S_H^T) = k$ this implies that $u^T H = v^T$, multiplying by A gives us the required result $u^T W = u^T HA^T = v^T A^T$. □

Claim A.3. *Let $\overline{M} = [0^{n\times n}|HS_H^T]$, then $u^T M = q'^T$ if and only if $u^T\overline{M} = q'^T$. Note that the columns of $\overline{M}$ are spanned by the columns of H which is rank k, whereas our assumption on M is that it is rank $2k$.*

Proof. In the course of arguing the previous claim, we showed that

$$u^T M = q'^T = [0^n|q^T] = [0^n|u^T HS_H^T] \ ,$$

but $u^T\overline{M} = [0^n|u^T HS_H^T]$ too. □

Claim A.4. *Let $u^T M = q'^T$, let $(\overline{M})^{-1}$ be the pseudo inverse of $\overline{M}$, and M^{-1} be the pseudo inverse of M, then $u^T = q'^T(\overline{M})^{-1} = q'^T M^{-1}$.*

Proof. Follows from Claim (A.3) and the fact that q'^T is in the row spaces of both $\overline{F}$ and F. □

Claim A.5. *Let $q^T = v^T S_H^T$, $q'^T = [0^n|q^T]$, $\overline{M} = [0^{k\times n}|HS_H^T]$, then*

$$v^T A^T = q'^T(\overline{M})^{-1}W = q'^T M^{-1}W.$$

Proof. Follows from Claims (A.2) and (A.4). □

Claim A.6. *Let $q^T = v^T S_H^T$, $q'^T = [0^n|q^T]$, then $(q'^T M^{-1})^T$ is in the column space of W.*

Proof. Recall that $\overline{M} = [0^{n\times n}|HS_H^T]$, the SVD of $\overline{M} = U_{\overline{M}}\Sigma_{\overline{M}}V_{\overline{M}}^T$. Given that $\overline{M}$ is of rank k, the columns of $U_{\overline{M}}$ span the same space as the columns of H. As $W = HA^T$ and W is rank k the columns of W and the columns of H span the same space. *I.e.*, the column space of $U_{\overline{M}}$ is the same as the column space of W.

From Claim A.3 we know that $q'^T M^{-1} = q'^T\overline{M}^{-1} = (q'^T V_{\overline{M}}\Sigma_{\overline{M}}^{-1})U_{\overline{M}}^T$. *I.e.*, $q'^T M^{-1}$ is a linear combination of the rows of $U_{\overline{M}}^T$, or alternately $(q'^T M^{-1})^T$ is in the column space of $U_{\overline{M}}$ which is the same as the column space of W. □

A.0.2 Lower bound on $\|v^T A^T\|_2$

We know from Claim A.5 that

$$\begin{aligned} v^T A^T &= q'^T M^{-1} W \\ &= q'^T V_M \Sigma_M^{-1} U_M^T U_W \Sigma_W V_W^T. \end{aligned} \tag{3}$$

We derive a lower bound on Equation 3 using the facts that:

1. q'^T is in the row space of M, which means that it's in the column space of V_M (Claim A.1);
2. for any $B \in R^{i\times j}$, $i \geq j$, whose columns have 2-norm 1 and are mutually orthogonal[11] and $z \in R^i$, a vector in the column space of B, $\|z^T B\|_2 = \|z^T\|_2$;
3. The smallest singular value of Σ_M^{-1} is $1/\sigma_1(M)$;
4. $(q'^T M^{-1})^T$ is in the column space of W (and thus of U_W) (Claim A.6); and
5. The smallest singular value of Σ_W is $\sigma_k(W)$,

yielding the following lemma.

Lemma A.7.

$$\|v^T A^T\|_2 \geq \|q'^T\|_2 \frac{\sigma_k(W)}{\sigma_1(M)}.$$

[11] We use this cumbersome definition because orthogonal matrices are by definition square.

A.0.3 Upper bound

We use the equality $v^T A^T = q'^T M^{-1} W$ from Claim A.5. Also, let $e \in R^{n+\ell}$ be such that $\widehat{q}'^T = q'^T + e$, let $E_{M^{-1}} \in R^{(n+\ell)\times n}$ be such that $\widehat{M}_m^{-1} = M^{-1} + E_{M^{-1}}$ and let $E_W \in R^{n\times n}$ be such that $\widehat{W}_r = W + E_W$. Then we can write:

$$\begin{aligned}
\| \quad \widehat{q}'^T \quad & \widehat{M}_m^{-1}\widehat{W}_r - v^T A^T\|_2 \\
= \quad & \|\widehat{q}'^T \widehat{M}_m^{-1}\widehat{W}_r - q'^T M^{-1} W\|_2 \\
= \quad & \|(q'^T + e^T)(M^{-1} \\
& \quad +E_{M^{-1}})(W + E_W) \\
& \quad -q'^T M^{-1} W\|_2 \\
\leq \quad & \|e^T M^{-1} W\|_2 && (4) \\
& +\|q'^T M^{-1} E_W + e^T M^{-1} E_W\|_2 && (5) \\
& +\|q'^T E_{M^{-1}} W + e^T E_{M_{-1}} W\|_2 && (6) \\
& +\|q'^T E_{M^{-1}} E_W + e^T E_{M^{-1}} E_W\|_2 && (7)
\end{aligned}$$

By a simple martingale argument omitted here (essentially Lemma 7.0.4 in [2]), we can show that:

Claim A.8. *For any fixed matrix $B \in R^{(n+\ell)\times j}$ with constant rank i, $\|e^T B\|_2 \leq O(1)\sqrt{i}\cdot\sigma_1(B)$ with high probability.*

Claim A.9. *If $\sigma_i(W) \in \omega(\sqrt{n})$ for $1 \leq i \leq k$, [12] then SP chooses $r = k$ and $\|E_W\|_2 \in O(\sqrt{n})$ with high probability [13]. Similarly, if $\sigma_i(M) \in \omega(\sqrt{n+\ell})$ for $1 \leq i \leq 2k$ then SP chooses $m = 2k$ and $\|E_M\|_2 \in O(\sqrt{n+\ell})$ with high probability, where E_M is the matrix such that $\widehat{M}_r = M + E_M$[14].*

Proof. We can write $\|E_W\|_2 = \|W - \widehat{W}_r\|_2 \leq \|W - \widehat{W}\|_2 + \|\widehat{W}_r - \widehat{W}\|_2$. Now, $E = W - \widehat{W}$ is a matrix of random variables with mean 0 and constant range, and the 2-norm of such a matrix is almost certainly $O(\sqrt{n})$[4].

We can also bound $\|\widehat{W}_r - \widehat{W}\|_2 \leq \sigma_{r+1}(\widehat{W})$. We now observe that for all i $\sigma_i(\widehat{W}) = \sigma_i(W - E) \leq \sigma_i(W) + \|E\|_2$.

We know that $\sigma_{k+1}(W) = 0$ and by assumption $\sigma_i(W) \in \omega(\sqrt{n})$ for all $1 \leq i \leq k$, therefore, $\sigma_k(\widehat{W}) \in \omega(\sqrt{n})$ and $\sigma_{k+1}(\widehat{W}) \in O(\sqrt{n})$, which means that *SP* chooses $r = k$. Thus, $\|\widehat{W}_r - \widehat{W}\|_2 \leq \sigma_{k+1}(\widehat{W}) \in O(\sqrt{n})$.

[12] By assumption, $\sigma_i(W) \in \omega(n^\delta)$ which is much higher and thus this condition clearly holds, we state the weaker requirement so as to explain what conditions are required for every step of the way. Likewise, the conditions of the successive claims hold by the assumptions in Theorem 3.1.

[13] The probability space here is defined by the random choices made when instantiating W to get $\widehat{W}$.

[14] The probability space is defined by the random choices made when instantiating W to get $\widehat{W}$ and the random choices made when instantiating S to get $\widehat{S}$.

Similar arguments can be used to show the other part of the claim. □

Claim A.10. *If $\sigma_i(M) \in \omega(\sqrt{n+\ell})$ for $1 \leq i \leq 2k$, then, with high probability*

$$\|E_{M^{-1}}\|_2 \leq \frac{O(\sqrt{n+\ell})}{(\sigma_{2k}(M))^2}.$$

Proof. By Wedin's theorem (see [27], — Theorem 3.8, p. 143) we get

$$\begin{aligned}
\|E_{M^{-1}}\|_2 &= \|\widehat{M}^{-1} - M^{-1}\|_2 \\
&\leq O(1)\max(\|M^{-1}\|_2^2, \|\widehat{M}_r^{-1}\|_2^2)\|E_M\|_2.
\end{aligned}$$

Now, $\|M^{-1}\|_2 = 1/\sigma_{2k}(M)$, and $\|\widehat{M}_r^{-1}\|_2 = 1/\sigma_{2k}(\widehat{M}_r)$. We also know that that $\sigma_{2k}(\widehat{M}_r) \leq \sigma_{2k}(M) + O(\|E_M\|_2) = \sigma_{2k}(M) + O(\sqrt{n+\ell})$, whereas by assumption $\sigma_{2k}(M) \in \omega(\sqrt{n+\ell})$ which implies that $\sigma_{2k}(\widehat{M}_r) = \Theta(\sigma_{2k}(M))$.

Thus,

$$\begin{aligned}
\|E_{M^{-1}}\|_2 &\leq O(1)\cdot\|E_M\|_2/(\sigma_{2k}(M))^2 \\
&\leq O(\sqrt{n+\ell})/\sigma_{2k}(M)^2,
\end{aligned}$$

by Claim A.9 □

Using Claims A.8, A.9, A.10, the facts that $\|M^{-1}\|_2 = 1/\sigma_{2k}(M)$ and $\|W\|_2 = \sigma_1(W)$, and some simple algebra, we easily obtain upper bounds on Equations (4), (5), (6) and (7). These upper bounds can then be combined with Lemma A.7, to show that if the length of the search query $\|q^T\|_2 \in \omega(\sqrt{k}\cdot r_k(W) r_{2k}(M))$, then

$$\frac{\|\widehat{q}'^T \widehat{M}_m^{-1}\widehat{W}_r - v^T A^T\|_2}{\|v^T A^T\|_2} = o(1),$$

which completes the proof of Theorem 3.1. □