Learning to Probabilistically Identify Authoritative Documents

David Cohn COHN@JUSTRESEARCH.COM
Just Research, 4616 Henry Street, Pittsburgh, PA 15213

Huan Chang HUAN@CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

We describe a model of document citation that learns to identify hubs and authorities in a set of linked documents, such as pages retrieved from the world wide web, or papers retrieved from a research paper archive. Unlike the popular HITS algorithm, which relies on dubious statistical assumptions, our model provides probabilistic estimates that have clear semantics. We also find that in general, the identified authoritative documents correspond better to human intuition.

1. Introduction

Bibliometrics has been described as a "series of techniques that seek to quantify the process of written communication" (Ikpaahindi, 1985). It typically attempts to give quantified answers to questions involving the relationships among documents, or authors and documents: "Who are the authoritative authors in this field?" "What are the seminal papers?" "How many distinct communities are studying this subject?" and others (see White & McCain, 1989, for details). Traditionally, the statistics upon which this quantification is based are citations in scientific literature; with the advent of the world wide web, it has become popular to apply bibliometric techniques to the hyperlinks of web pages (Kleinberg, 1997; Larson, 1996) or even web page browsing patterns (Turnbull, 1996).

Recent applications to the web have caused a resurgence of interest in bibliometrics, especially when used in conjunction with information retrieval. Information retrieval is primarily concerned with identifying the "most relevant" document for a query; with the explosion in the size of the web however, users are often swamped by thousands of "equally relevant" pages. The challenge now faced by search engines is finding the pages that are most relevant *and* authoritative.

1.1 HITS and PCA

Although Larson (1996) pioneered the application of bibliometrics to the web, the most well known and popular bibliometric algorithm for the web is Kleinberg's "Hypertext-Induced Topic Selection" (HITS) algorithm and its variations (Bharat & Henzinger, 1998; Kleinberg, 1997). The process begins with a matrix M of document-citation pairs. Entry M_{ij} is nonzero iff document i cites document j or, equivalently, if i contains a hyperlink to j.

Traditionally, one generates the co-citation matrix $M'M$ or bibliographic coupling matrix MM' and attempts to identify correlations. These correlations are identified in the form of principal components (or eigenvectors) of the matrix, each of which corresponds to a "community" of roughly similar citation patterns.

HITS uses an iterative process (Golub & Loan, 1989) to identify the principal eigenvector (and principal community) of the matrix. The extent of this vector in a document's dimension is called the "loading" of the document on the vector. The loading of a document on the principal eigenvector of $M'M$ is interpreted as the "authority" of that document within the community — how likely it is to by cited within that community. Document loading on the principal eigenvector of MM' is interpreted as its "hub" value in the community — how many authoritative documents it cites within the community.

Because only the largest eigenvector is extracted, all but the principal community are ignored. It is possible for other, only slightly smaller communities to be skipped over, giving their "authoritative" documents no credit for their authority (see Figure 1). This problem is endemic in Information Retrieval, where synonyms in query or answer documents may cause retrieval of documents from multiple, unrelated topics.

This problem can be easily corrected. More traditional bibliometric methodology often employ principal com-

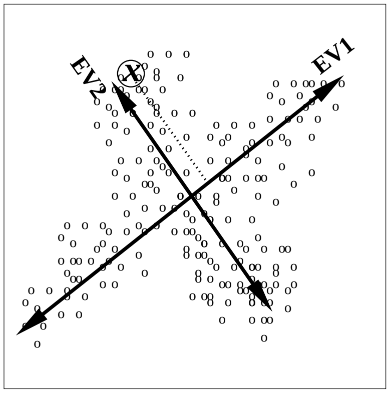

Figure 1. A weakness of the HITS algorithm is its focus on the single largest eigenvector when conveying authority. Here, eigenvectors EV1 and EV2 correspond to two orthogonal communities identified by co-citation factor analysis. Document x has great authority in in community EV2, by dint of its projection onto that axis. The community represented by EV1 is slightly larger however, so EV1 becomes the principal eigenvector. The projection of x onto EV1 is minimal, so it will be given little authority by HITS, despite the authority it commands along EV2.

ponents analysis (PCA) to extract multiple eigenvectors. PCA is a form of linear factor analysis: each of the eigenvectors serves as a factor that can be combined linearly with others and blurred with noise to "explain" variations in the data (R. Gorsuch, 1983). These multiple factors (eigenvalues) correspond to the largest bibliographic communities in $M'M$ and MM': documents that are heavily loaded on them command authority or hub value within their community.

1.2 Statistical Assumptions of PCA

Despite the attractiveness and simplicity of PCA and the HITS algorithm, they have a serious shortcoming: they are built on a faulty statistical foundation. PCA and HITS minimize the *distortion* of the factored approximation to MM' — the mean squared difference between observed and approximated values. This corresponds to making the assumption that all random variation in citation rates is due to Gaussian noise.

Unfortunately, the Gaussian is particulary ill-suited to modelling citation rates: it is symmetric about its mean and is able to generate citation frequencies that are less than zero, or arbitrarily large. A noise model which is better suited to modelling citation counts is the multinomial distribution. In this paper, we introduce a bibliometric method for identifying hub and authority documents based on this probabilistic model of citation. Mathematically, the model is almost identical to PLSI (Hofmann, 1999), which provides a probabilistic interpretation of term-document relationships.

2. A Latent Variable Model of Citation

Our model of citation is based on a two-way factor analysis that is, in most respects, identical to the "aspect model" used by Hofmann (1999) and others. Hofmann's PLSI is a probabilistic analogue of Latent Semantic Indexing (LSI); we therefore call our approach "PHITS," as it is a probabilistic analogue of the HITS algorithm.

The model attempts to explain two sets of observables (in our case *documents* and *citations*) in terms of a small number of common, but unobserved variables (also called aspects or factors). In bibliometrics, these factors are typically identified with individual research area, or "communities".

Statistically, we can describe the model as a generative process, borrowing notation from Hofmann (see Figure 2). A document $d \in D$ is generated with some probability $P(d)$. The factor, or topic $z \in Z$ associated with d is chosen probabilistically according to $P(z|d)$. Given the factor, citations $c \in C$ are generated probabilistically according to $P(c|z)$.[1]

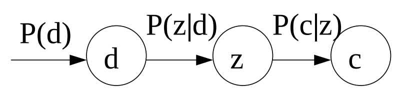

Figure 2. Factored statistical model of document citation.

Given a set of document-citation pairs (d, c), we can describe the likelihood of each pair as

$$\begin{aligned} P(d,c) &= P(d)P(c|d), \text{ where} \qquad (1)\\ P(c|d) &= \sum_z P(c|z)P(z|d), \end{aligned}$$

summing over all factors which could have produced the citation. The total likelihood L of the observed citations matrix M is then described as

$$L(M) = \prod_{(d,c)\in M} P(d,c). \qquad (2)$$

The process of building a model that "explains" a set of observations then reduces to the problem of finding values for $P(d)$, $P(z|d)$ and $P(c|z)$ that maximize the likelihood $L(M)$ of the observed data.

[1] In practice, C and D both refer to documents in the corpus, and the sets may be identical. We keep them separate notationally to reinforce different roles they play in the model: membership in C is conveyed by being cited, membership in D is conveyed by citing.

2.1 Mixture Models vs. Factored Models

It is important to briefly distinguish the factored model used here from probabilistic mixture models. In a mixture model, each example is assumed to come from *one* of a set of latent sources (e.g., a document is either about z_1 or z_2). Credit for the example may be distributed among several sources because of ambiguity, but the model insists that only one of the candidate sources is the *true* origin of the example. In contrast, the factored model assumes that all examples come from a *combination* of sources — without any ambiguity, the model can assert that a document is half z_1 and half z_2.

2.2 Maximizing Model Likelihood

To find a maximally likely model, we again borrow from Hofmann, using a "tempered" version of the EM algorithm (Hofmann, 1999). We make use of Bayes rule to reformulate Equation 1 in terms of the latent variable z:

$$P(d,c) = \sum_z P(z)P(c|z)P(d|z). \tag{3}$$

We then alternate steps of computing expectations of $P(z|d,c)$ with steps of re-estimating $P(z)$, $P(c|z)$ and $P(d|z)$ to maximizing the data likelihood. Beginning with an arbitrary values of $P(z)$, $P(c|z)$ and $P(d|z)$, and given some fairly gentle assumptions, this iteration is guaranteed to converge to a locally optimal likelihood solution (Dempster et al., 1977). Unlike PCA, however, there are no guarantees on the global optimality of the resulting solution.

The expectation step consists of computing

$$P(z|d,c) = \frac{P(z)P(d|z)P(c|z)}{\sum_{z'} P(z')P(d|z')P(c|z')}, \tag{4}$$

for each z, d, and c — the expectation that a particular document-citation pair is "explained" by factor z.

Using these values, new maximum likelihood estimates are derived for the conditional likelihoods of the observables:

$$P(z) = \frac{\sum_{d,c} P(z|d,c)}{\sum_{z'}\sum_{d,c} P(z'|d,c)} \tag{5}$$

$$P(d|z) = \frac{\sum_{c} P(z|d,c)}{\sum_{d',c} P(z|d',c)} \tag{6}$$

$$P(c|z) = \frac{\sum_{d} P(z|d,c)}{\sum_{d,c'} P(z|d,c')} \tag{7}$$

Repeated computation of Equation 4 interleaved with Equations 5–7 monotonically increases the total likelihood of the observed data $L(M)$.

In addition to the basic EM algorithm, we also apply Hofmann's tempering variation: we "temper" the assignment of factors with a parameter β by replacing Equation 4 with

$$P(z|d,c) = \frac{[P(z)P(d|z)P(c|z)]^\beta}{\sum_{z'} [P(z')P(d|z')P(c|z')]^\beta}. \tag{8}$$

We run EM beginning with a value $\beta = 1$, iterate until the data likelihood improvement is negligible, then slowly reduce β and repeat (see Hofmann, 1999, for details).[2] We use a hard lower limit on β; the best value to use for that limit appears to vary with the size and connectivity of the corpus.

3. Experiments

We test our model on two corpora. The first is in the traditional bibliometric domain of citations among scientific papers; the second is on a set of hyperlinked documents retrieved from the world wide web.

3.1 Citation Communities in Cora

Cora (McCallum et al., 1999; McCallum et al., 2000) is an online archive of computer science research papers. The archive was built automatically using a combination of smart spidering to efficiently find online papers in PostScript format, information extraction to identify paper titles, authors, abstracts and references, and statistical text classification to categorize the papers into a Yahoo-like topic hierarchy with approximately 70 leaf categories. The archive contains approximately 30,000 papers, and over 1 million links to roughly 200,000 distinct documents.

We can use the subtopic categorization of *Cora* as a form of objective test set. Since *Cora* classifies papers according to text content, its classifications are independent of the citation patterns between documents. If the research areas identified by *Cora*'s classifications do indeed represent distinct communities (which may or may not cite each other), we can hope that our algorithm can recover these communities from citation patterns, and identify the most influential papers within each community.

To test this hypothesis, we selected a subset of the *Cora* database, the papers under the "Machine Learning" category. This category is subdivided into 7 subtopics: Case-Based Reasoning, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning and Theory. We

[2] Data likelihood is measured using a form of leave-one-out validation, where each $P(z|d,c)$ is evaluated without counting the occurrence of the (d,c) pair in question.

identified each of the 4240 documents and 38,372 citations in these subtopics and used them as training data for the PHITS algorithm.

We asked the system to extract 7 factors from the data, and seeded the model randomly by setting $P(d|z)$: each document received uniform membership in all factors except one chosen at random, which received "double" weight. The citations from each document were given a corresponding distribution in $P(c|z)$, and $P(z)$ was set to be uniform. We began with a tempering value of $\beta = 1$ and decreased it by a multiple of 0.9 every 20 iterations, or whenever the improvement in model likelihood fell below $0.00001L$. We terminated the algorithm after a maximum of 40 iterations for efficiency — although likelihood generally continued to increase beyond this point, we found that the increase had little effect on the factors or orderings of the most probable citations.

3.1.1 Interpreting the Results

Kleinberg's HITS algorithm grants authority to a document proportional to the magnitude of its component in the largest eigenvector of the co-citation matrix. While indicative of the influence that document has on others aligned with the eigenvector, no probabilistic interpretation is possible using this model.

Given a fully probabilistic model, there are a number of similar statistics we could use to measure the importance of a document given a factor (i.e. within a community of citations). The simplest analog to "authority" is the conditional probability $P(c|z)$. This indicates how likely a document c is to be cited from within community z. Table 1 lists the documents with highest $P(c|z)$ for each of the 7 learned factors. The PHITS recovery of the original categories is remarkably accurate; more importantly, several generally recognized authoritative papers for each category appear at the top.

The traditional HITS algorithm also extracts some authoritative papers for each of its categories, but there is less of a clear distinction between categories, and the list contains a number of papers that are not generally considered authorities. Interestingly, HITS identifies as authoritative a category which does not appear explicitly in Cora: classical statistics.

3.1.2 Alternative Statistics

It is worth noting that $P(c|z)$ is only one of the statistics which may be computed with the probabilistic model. It corresponds to how authoritative a document is considered to be from within a community. This is entirely distinct from the question of which community the document itself is in — a seminal theory paper, for example, may be authoritative in many other fields. To find community membership, we simply compute

$$P(z|c) = \frac{P(c|z)P(z)}{P(c)} = \frac{P(c|z)P(z)}{\sum_{z'} P(c|z')}. \qquad (9)$$

according to Bayes rule. This quantity can be used to classify documents according to communities. For example, the paper "TD Learning of Game Evaluation Functions with Hierarchical Neural Architectures," by Wiering (1995), has factor probabilities

0.566	*Reinforcement Learning*
0.027	*Rule Learning*
0.239	*Neural Networks*
0.026	*Theory*
0.026	*Probabilistic Reasoning*
0.072	*Genetic Algorithms*
0.044	*Logic*

indicating that as we would expect, it is primarily a mix of reinforcement learning and neural networks, with marginal membership in other communities.

Beyond classification, there is another interesting use of $P(z|c)$ – by examining the authority of documents which have one dominating factor, we can identify papers which are topic-specific authorities. We can, for example, look at which theory papers are most authoritative with respect to the Neural Network (NN) community by examining

$$argmax_c P(c|z = NN) : P(z = Theory|c) > 0.9. \qquad (10)$$

The paper which maximizes this quantity is "Decision Theoretic Generalizations of the PAC Model for Neural Net and other Learning Applications," by Haussler (1992).

Table 1. (top) Highest ranked documents in Machine Learning, according to PHITS' computation of $p(c|z)$, the probability of a citation c being referenced from a document in this factor. Labels in parentheses were attached manually, corresponding to one of the 7 original topic labels from Cora. (bottom) Highest ranked documents in Machine Learning, according to HITS' eigenvector computation. Labels in parentheses were attached manually, corresponding when possible to topic labels from Cora – "Classical Statistics" is a category not explicitly represented in Cora.

Top citations by $P(c|z)$, computed by PHITS algorithm:	
factor 1	(Reinforcement Learning)
0.0108	Learning to predict by the methods of temporal differences. Sutton
0.0066	Neuronlike adaptive elements that can solve difficult learning control problems. Barto et al
0.0065	Practical Issues in Temporal Difference Learning. Tesauro.
factor 2	(Rule Learning)
0.0038	Explanation-based generalization: a unifying view. Mitchell et al
0.0037	Learning internal representations by error propagation. Rumelhart et al
0.0036	Explanation-Based Learning: An Alternative View. DeJong et al
factor 3	(Neural Networks)
0.0120	Learning internal representations by error propagation. Rumelhart et al
0.0061	Neural networks and the bias-variance dilemma. Geman et al
0.0049	The Cascade-Correlation learning architecture. Fahlman et al
factor 4	(Theory)
0.0093	Classification and Regression Trees. Breiman et al
0.0066	Learnability and the Vapnik-Chervonenkis dimension, Blumer et al
0.0055	Learning Quickly when Irrelevant Attributes Abound. Littlestone
factor 5	(Probabilistic Reasoning)
0.0118	Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Pearl.
0.0094	Maximum likelihood from incomplete data via the em algorithm. Dempster et al
0.0056	Local computations with probabilities on graphical structures... Lauritzen et al
factor 6	(Genetic Algorithms)
0.0157	Genetic Algorithms in Search, Optimization, and Machine Learning. Goldberg
0.0132	Adaptation in Natural and Artificial Systems. Holland
0.0096	Genetic Programming: On the Programming of Computers by Means of Natural Selection. Koza
factor 7	(Logic)
0.0063	Efficient induction of logic programs. Muggleton et al
0.0054	Learning logical definitions from relations. Quinlan.
0.0033	Inductive Logic Programming Techniques and Applications. Lavrac et al

Top citations by eigenvector:	
factor 1	(Genetic Algorithms)
0.0492	How genetic algorithms work: A critical look at implicit parallelism. Grefenstette
0.0490	A theory and methodology of inductive learning. Michalski
0.0473	Co-evolving parasites improve simulated evolution as an optimization procedure. Hills
factor 2	(Genetic Algorithms)
0.00295	Induction of finite automata by genetic algorithms. Zhou et al
0.00295	Implementation of massively parallel genetic algorithm on the MasPar MP-1. Logar et al
0.00294	Genetic programming: A new paradigm for control and analysis. Hampo
factor 3	(Reinforcement Learning/Genetic Algorithms)
0.256	Learning to predict by the methods of temporal differences. Sutton
0.238	Genetic Algorithms in Search, Optimization, and Machine Learning. Angeline et al
0.178	Adaptation in Natural and Artificial Systems. Holland
factor 4	(Neural Networks)
0.162	Learning internal representations by error propagation. Rumelhart et al
0.129	Pattern Recognition and Neural Networks. Lawrence et al
0.127	Self-Organization and Associative Memory. Hasselmo et al
factor 5	(Rule Learning)
0.0828	Irrelevant features and the subset selection problem, Cohen et al
0.0721	Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Holte
0.0680	Classification and Regression Trees. Breiman et al
factor 6	(Rule Learning)
0.130	Classification and Regression Trees. Breiman et al
0.0879	The CN2 induction algorithm. Clark et al
0.0751	Boolean Feature Discovery in Empirical Learning. Pagallo
factor 7	([Classical Statistics?])
1.5-132	Method of Least Squares. Gauss
1.5-132	The historical development of the Gauss linear model. Seal
1.5-132	A Treatise on the Adjustment of Observations. Wright

Another measure we might want is an indication of what documents are "characteristic" of a community. This concept is distinct from both membership and raw probability: a poorly-cited document that has exclusive membership in one community is not characteristic of that community; neither is a heavily-cited document, if it is equally heavily-cited by all other communities. Identifying heavily-cited documents that are community-specific is a matter of computing the product of the terms: $char_{ij} = P(z_j|c_i) \cdot P(c_i|z_j)$.

The probabilistic equivalent of "hubs" are found via the same route. $P(d|z)$ is the probability, given some factor z, that document d contains a reference to it. The more factor-specific references d makes, the greater $P(d|z)$ will be. As such, this probability serves the same function as a hub score. Just as with citations, we can compute the membership $P(z|d)$ of a document, or its characteristic probability $P(z|d) \cdot P(d|z)$.

3.2 Authoritative Web Documents

The primary domain considered by Kleinberg (1997) was hyperlinked documents on the world wide web. Indeed, as the number web pages "relevant" to a given search engine query has exploded, the need for a means of identifying authoritative pages has grown more crucial. Comparing the performance of HITS and PHITS in this domain reveals some interesting contrasts.

We began by gathering a set of documents and links following the HITS methodology:

1. Issue a query to Altavista (`www.altavista.com`) and define the pages it returns as a "root set."
2. Add to the root set all pages with links pointing to the root set, and all pages pointed to by the root set. This defines the "base set."
3. Remove "intrinsic" links — links between two sites sharing the same top level domain. Keep all other links between members of the base set.

The base set and its remaining hyperlinks are the documents and citations to which we apply PHITS.

Here, we describe results for the "jaguar" query considered in Kleinberg's paper; we have obtained similar results with other queries described in that paper.[3] Issuing an Altavista query on the terms "jaguar" and "jaguars" yielded a base set of 5276 links between 2372 web pages.

In Kleinberg's results, the dominating communities were for the Atari Jaguar product, the Jaguars football team, and the Jaguar automobile. Oddly, the principal eigenvector (magnitude 729) returned by HITS involves services in Cincinnati – dominated by a clique of sites from a single domain:

Principal eigenvector (magnitude 729.84)	
0.224	http://www.gannett.com
0.224	http://homefinder.cincinnati.com
0.224	http://cincinnati.com/freetime/movies
0.224	http://autofinder.cincinnati.com

The cause of this behavior is the Cincinnati Enquirer newspaper, which has many articles about the Cincinnati Bengals and their rival, the Jacksonville Jaguars. Each article contains the same set of pointers to services provided by the newspaper, leading these to dominate the link set.

The second eigenvector is dominated by the home pages of technology news agencies — authoritative in their fields, but not obviously associated with jaguars:

Third eigenvector (magnitude 358.39)	
0.0003	http://www.cmpnet.com
0.0003	http://www.networkcomputing.com
0.0002	http://www.techweb.com/news
0.0002	http://www.byte.com

In fact, it is not until the third eigenvector (magnitude 294) that the Jaguars football team appears:

Third eigenvector (magnitude 294.25)	
0.781	http://www.jaguarsnfl.com
0.381	http://www.nfl.com
0.343	http://jaguars.jacksonville.com
0.174	http://www.nfl.com/jaguars

The Jaguar automobile appears at the negative end of the fourth eigenvector (magnitude 169), and the Atari product is nowhere to be found (a victim of progress).

The same data, run on PHITS, exhibits very different behavior. Run with two factors (which should capture the two largest components), we find that the Jacksonville Jaguars and the Jaguar automobile do dominate the two factors. However, unlike HITS, these two topics are not segregated into one factor each, but interleaved in both:

Top citations by $P(c|z)$:	
Factor 1	
0.0440	http://www.jaguarsnfl.com -
0.0252	http://jaguars.jacksonville.com -
0.0232	http://www.jag-lovers.org -
0.0200	http://www.nfl.com -
0.0167	http://www.jaguarcars.com -
Factor 2	
0.0367	http://www.jaguarsnfl.com -
0.0233	http://www.jag-lovers.org -
0.0210	http://jaguars.jacksonville.com -
0.0201	http://www.nfl.com -
0.0161	http://www.jaguarcars.com -

[3] The original link sets used by Kleinberg are not publicly available, and the topology of the web has changed substantially in the three years since those experiments, so it is not surprising that our results differ.

Running with more or fewer factors yields a similar result: while we identify what are arguably the most authoritative pages returned by the query, PHITS has not separated them into distinct factors. This highlights a distinction between the PCA-based HITS algorithm and ours. PCA forms orthogonal eigenvectors. Probabilistically, there is no requirement that citations belong to exclusively one factor. In the above example, PHITS appears to have found that it can maximize the likelihood of the observed data by adopting a shared representation.

In this case, our complaint should not be with the factoring, but with our choice of a factored model over a mixture model (Section 2.1). In the case of the "jaguars" query, we would expect each page to be about *one* of the four major categories — we don't actually expect to see any web pages that are a 50%/50% mix of British cars and Atari computers. As such, a traditional mixture model would be appropriate for first clustering the disjoint topics, after which PHITS could be applied to tease apart the different factors and authoritative web pages in each cluster.

There is another alternative, if we insist on orthogonality in the factored model. We could augment the the likelihood maximization process with constraints leading to an independent components analysis (Bell & Sejnowski, 1996) consistent with the probabilistic model. These approaches are the subject of work in progress.

4. Discussion

The factored probabilistic model of citations described here has both advantages and disadvantages with respect to the traditional linear factorization. The most important advantage is that, since it is based on probabilities, the estimates it provides have a clear interpretation: where the traditional model provides scalar magnitudes of authority, our model estimates actual probabilities. These probabilities have well-understood semantics, and may be combined and manipulated to provide answers to quantitative bibliometric questions.

Subjectively, the authorities extracted by our probabilistic model seem to correspond better to human intuition than those extracted by PCA and the HITS model. On citation data, it robustly recovers established categories much better than PCA. On web links, it recovers the intuitive communities, but frequently mixes their representation. Part of difference may be due to mismatch in the statistical model, but another factor may be that citations are more "intentional" than web links — `www.microsoft.com/ie` appears uniformly authoritative across all topics on the web because so many pages carry the linked banner "This page best viewed with Internet Explorer." Aside from enforcing factor orthogonality, a means of weighting the authority of a page by its relevance (determined by word content) should alleviate these spurious associations. See the end of this section for our current work in this area.

Beyond unintended mixing of communities, there are several minor disadvantages to the current probabilistic model, most of which are the subject for future research. First, unlike the eigenvector model derived from PCA, the EM-based training of PHITS is not guaranteed to identify a globally "optimal" factoring. The performance of the EM procedure is dependent on the starting point of the optimization, and may get stuck in local optima with poor overall performance.

Empirically, we observe some variation in the factors and corresponding authorities between runs, but the primary trends are fairly constant. We have not observed noticeably "bad" fits in any of our experiments, but we have no guarantee that we're not missing a "perfect" global fit somewhere. Other than performing multiple restarts on the fitting procedure, there is little that can be done to address this shortcoming. Another interesting possibility is to begin by computing the PCA-based model, and use those factors to seed the probabilistic model.

Another shortcoming of the probabilistic model is that, in the present implementation, we must decide *a priori* on the number of factors to model. PCA permits extracting successive factors iteratively and observing their magnitude. By looking for a tapering-off in the magnitude of the extracted factors (called a "scree test"), one can estimate when all significant factors have been extracted.

Roughly the same may be accomplished with a probabilistic model, although the process trades computational expense for the risk of getting stuck in local maxima. Given a factored probabilistic model, one can select a factor to "split" into subfactors. The subfactors can be re-fit, and the resulting increase in model likelihood examined. Using a model selection criterion such as AIC (Akaike, 1990) or BIC (Schwarz, 1978), one can determine whether the increase in likelihood justifies the split. If so, the split is kept, another factor is selected for splitting, and the process is repeated. The danger here, as with all hierarchical techniques, is that an early split which appears optimal may eventually lead to suboptimal splits later on. Automatically identifying the "right" number of factors, and learning a hierarchy of increasingly specific factors are two extensions of this work that we are exploring.

Computationally, the approaches are comparable. Running in MATLAB with sparse matrix routines on a Pentium III-550, HITS required approximately 5 seconds to compute anywhere between 2–5 factors. On the same machine, PHITS, required 4 seconds to compute 2 factors, and 12 seconds to compute 5 factors.

A final advantage of the probabilistic factored model over the traditional one is that it provides a foundation for building a unified probabilistic model of the content and connections of linked documents. Hofmann's PLSI performs a two way term-document factoring. Our model performs a two way document-citation factoring. It is mathematically straightforward to combine these models into a single three way factoring that relates terms, citations and documents in a unified probabilistic framework. We are currently developing a system which implements this factoring, and hope to soon report results from its use.

References

Akaike, H. (1990). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723.

Bell, A., & Sejnowski, T. (1996). An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillations in a hyperlinked environment,. *Twenty First ACM-SIGIR Conference on Information Retrieval.*

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society Series B*, *39*, 1–38.

Golub, G., & Loan, C. V. (1989). *Matrix computations.* Johns Hopkins University Press.

Haussler, D. (1992). Decision theoretic generations of the pac-model for neural nets and other applications. *Information and Computation*, *100*, 78–150.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Twenty Second International ACM-SIGIR Conference on Information Retrieval.*

Ikpaahindi, L. (1985). An overview of bibliometrics: its measurements, laws, and their applications. *Librarian*, 163–176.

Kleinberg, J. (1997). Authoritative sources in a hyperlinked environment. *Proceedings of the ACM–SIAM Symposium on Discrete Algorithms.*

Larson, R. R. (1996). *Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace* (Technical Report TRCS96-05). Computer Science Department, University of California, Santa Barbara.

McCallum, A., Nigam, K., Reed, J., Rennie, J., & Seymore, K. (2000). *Cora, a computer science research archive* (Technical Report). Just Research, http://www.cora.justresearch.com/.

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (1999). Building domain-specific search engines with machine learning techniques. *Working Notes of the AAAI Spring Symposium on Intelligent Agents in Cyberspace.*

R. Gorsuch, R. (1983). *Factor analysis.* Lawrence Erlbaum Associates.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Turnbull, D. (1996). *Bibliometrics and the web* (Technical Report FIS-12-19-1996-1). Faculty of Information Studies, University of Toronto.

White, H., & McCain, W. (1989). Bibliometrics. *Annual Review of Information Science Technology*, *24*, 119–165.

Wiering, M. (1995). *Td learning of game evaluation functions with hierarchical neural architectures.* Doctoral dissertation, Department of Mathematics and Computer Science University of Amsterdam.