Learning to Create Customized Authority Lists

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

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

Abstract

The proliferation of hypertext and the popularity of Kleinberg's HITS algorithm have brought about an increased interest in link analysis. While HITS and its older relatives from the Bibliometrics provide a method for finding authoritative sources on a particular topic, they do not allow individual users to inject their own opinions on what sources are authoritative. This paper presents a technique for learning a user's internal model of authority. We present experimental results based on *Cora* on-line index, a database of approximately one million on-line computer science literature references.

1. Introduction

Bibliometrics (White & McCain, 1989; Small, 1973) involves studying the structure that emerges from sets of linked documents. Traditionally, these links have taken the form of citations among journal articles, although Kleinberg (1997) and others (e.g., Brin & Page, 1998) have found that they adapt well to sets of hyperlinked documents. Bibliometric techniques exist for identifying subject areas, research specialties, and influential contributions to a corpus by examining correlations in the references or links between documents in the corpus.

Kleinberg's algorithm, Hypertext Induced Topic Selection (HITS), calculates principal eigenvector of the document link matrix, and ranks the authority of a document by the magnitude of its projection onto these eigenvectors. A deficiency of this approach is that, while the most popular (and thus most heavily-linked) documents may be authoritative in many senses, they may not correspond to a particular user's internal model of "authority." A user may have the opinion, for example, that a particular document X is truly an authority, or that anything written by author Y is worthless. What this user would like is a measure of document authority that coincides with their own preconceived notions of what is important.

In this paper, we present a technique that learns from a small amount of user feedback to realign the eigenvectors of the link matrix, effectively re-calibrating the measure of authority to correspond more closely to the user's internal model. Our technique is similar to relevance feedback (van Rijsbergen, 1979), but instead of manipulating a query to learn the *relevance* of the retrieved documents, we manipulate the weighting of the link matrix to learn their *authority*.

We demonstrate our algorithm on the problem of identifying subjectively authoritative computer science research papers. We first demonstrate how it can "lift" the authority of research papers in a particular subdiscipline, and then show how it can be used to automatically create user-specific "top ten" lists.

2. The "Authority" of a Document

In this section, we first describe how the "authority" of a document is computed with respect to a corpus, following the conventions of the HITS algorithm. We then describe the application of this method to a document corpus and discuss some of the limitations we have observed with its use. The following section describes our "lifting" algorithm, which addresses some of these limitations.

Hypertext-induced topic selection is performed as follows: Given a set of documents, a link matrix M specifies their connectivity: element M_{ij} is non-zero if and only if there is a reference or link from document i to document j. Typically M_{ij} is set to 1, or in some variations over the total number of references the document makes. For a given document i in the set, let a_i and

h_i be the "authority" and "hub" scores, respectively. These scores are real numbers greater than or equal to zero, and have the following interpretation: a large hub score means the document points to many good authorities; a large authority score means it is pointed to by many good hubs. This recursive definition leads to a set of linear equations

$$a_i = \sum_j M_{ji} h_j, \qquad h_i = \sum_j M_{ij} a_j \tag{1}$$

whose solutions are the corresponding eigenvectors of $M'M$ and MM' respectively. By defining vectors $a = [a_1, a_2, \ldots a_n]'$ and $h = [h_1, h_2, \ldots h_n]'$, we can express the equations in compact matrix form:

$$a = M'h = M'Ma, \qquad h = Ma = MM'h. \tag{2}$$

Rather than explicitly solving the system, it is often more efficient to pursue an iterative solution beginning with arbitrary values for a or h (usually the unit vector), and repeatedly applying Equation 2, normalizing a on each iteration by dividing each of its elements by $\sum_i a_i$. Subject to some mild assumptions, a is guaranteed to converge to the principal eigenvector of $M'M$ and h is guaranteed to converge to the principal eigenvector of MM' (Golub & Loan, 1989; Kleinberg, 1997).

Kleinberg (1997) reports results for experiments performed on web pages returned from a search engine query. Searching for the keyword "jaguar," he analyzes the set of documents returned in terms of their connectivity. He observes that the principal eigenvector of $M'M$ is well-aligned with authoritative pages about Atari Jaguar products. He notes, however, that the non-principal eigenvectors are populated by authoritative pages concerning other senses of the word "jaguar," such as the Jaguar baseball team, and the Jaguar automobile. This should come as no surprise: extracting the principal components of the matrix amounts to a form of linear factor analysis, so each principal component of $M'M$ represents an orthogonal direction in which there exist highly-correlated pages. "Communities" of closely-linked topics correspond exactly to orthogonal factors of the co-citation matrix, with the most heavily-linked of these communities emerging as the principal, or "authoritative" factor.

3. Experimental Domain: Citations from Cora

In order to evaluate our algorithm, we require a link matrix on which to operate. In the spirit of the Bibliometrics, we chose to use a matrix based on citation links in academic literature. One advantage of this domain over hyper-links in web pages is that authors think deliberately about creating reference links and are used to thinking carefully when interpreting the links. Furthermore, people have strong and differing personal opinions about what papers are truly authoritative, giving us a chance to test our "personally-customized authority lists."

Our link matrix consists of about one million references collected as part of the *Cora* index of computer science literature (*www.cora.justresearch.com*) (McCallum et al., 1999). Cora's smart spiders efficiently crawl the web in search of computer science papers in postscript format and turn the postscript into plain text. Statistical text classification is used to categorize the papers into a 70-leaf Yahoo-like topic hierarchy. The reference section of each paper is found with a hand-built regular expression, and the individual references are segmented by paragraph boundaries. Then a hidden Markov model, whose structure and parameters have been learned from training data, extracts the authors, title, institution, venue, etc., from each reference. This extraction is 93% accurate. Similar fields are extracted from the header of each paper. Finally references are automatically matched to papers and to each other using a soft-matching criteria that can detect when two references are referring to the same paper, even when the reference strings are not identical. The scheme uses a combination of a cheap distance metric based on TFIDF cosine-similarity and on a more powerful but expensive distance metric based on per-field string edit distance (McCallum et al., in press). Reference matching is accurate, but not perfect. On rare occasion, spurious links are created, and some links are missing which should be present.

This paper describes two types of experiments. In the first, the results of our modified HITS algorithm are compared to the results of the papers' *Cora* categorization into topic. In the second, we performed a user-study to measure the extent to which we could indeed create individually-customized authority lists.

3.1 Application to Cora and Limitations of HITS

Beginning with a target topic, or the set of documents retrieved by a query to *Cora*, it is relatively straightforward to apply the HITS algorithm. Following Kleinberg's convention, we begin with a root set of documents in the topic, and expand it to a base set of all documents that cite or are cited by documents in the root set before searching for the principal eigenvectors of their co-citation matrix.

In addition to identifying hub and authority papers, we ran a version of the algorithm on a link matrix consisting of author-to-author citations, in an attempt to identify authoritative authors. In this version, author A is linked to author B if A is an author of a paper that cites a paper containing B as an author.

The results from a naive application of the algorithm are mixed for both papers and authors. As we expected, hub scores tend to be a fairly good predictor of survey papers. A cursory inspection of the rankings reveals that papers with large hub scores often have titles such as "A Survey of...," "A Tutorial on...," "An Introduction to...," "A Guide to...," etc.

Evaluation of paper authority is less convincing. Papers that seem obscure to seasoned researchers in a field are often given higher ranks than well-recognized and heavily cited authorities. This sometimes-strange behavior has been noted elsewhere, and numerous attempts at a fix have been made (e.g., Bharat & Henzinger, 1998; Chakrabarti et al., 1998).

But the most severe problem with HITS ranking on our setup stems from the existence of secondary factors (or "communities," using Kleinberg's terminology) away from the principal eigenvector. Consider, for example, the leaf category called "Compression." Ranking by HITS authority puts 48 image and video compression research papers in the top 50. The 46^{th} and 50^{th} are papers on text compression. *Cora* classifies text and video compression together, but papers on text compression are less numerous, and the two communities do not cite each other, so text compression is shunted off to a secondary eigenvector. We can find these non-principle vectors by performing an singular value decomposition (SVD) on the citation matrix, but this process is impractical for large collections. Even if we do isolate all orthogonal communities, the community-dependent ranks are incomparable – there is no principled method for recombining ranks from the individual communities to identify what qualifies as a generally good authority on compression.

Humans seem to have little trouble identifying what they subjectively feel is an authority on the basis of personal knowledge. Often these authorities do not correspond with any of the principal eigenvectors of a co-citation matrix – nonetheless, they correspond with the human's internal model of authority. But having each user to manually rank the authority of all documents in a collection defeats the purpose of the task. What we would like, is to learn this internal model from the citation structure of the corpus–a lot of information provided by human authors, and a little commentary provided by a specific human user.

4. "Lifting" the Authority of a Document

We wish to find a way to raise the authority of documents that we personally consider important. More than that, we want the system to generalize from our actions, not only raising the authority of the page in question, but adjusting the authority of similar or related documents as appropriate. One approach would be to simply grant the paper in question the highest authority, maintaining it at this authority throughout HITS iterations, and rely on the spreading activation of the HITS algorithm to distribute that authority. But this approach suffers from nepotism: it only distributes authority according to how well and closely linked a document is to the raised document, without regard to the geometry of the co-citation space the documents are embedded in (see Figure 1).

We present another approach: instead of relying on spreading activation to directly confer authority, we perform gradient ascent on the elements of the link matrix, altering link weights to more closely align its principal eigenvector with the document(s) in question. The procedure by which we accomplish this is straightforward:

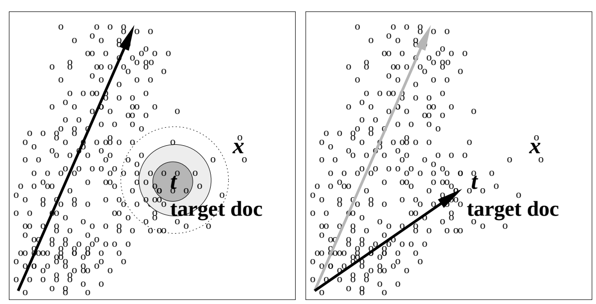

Figure 1. Illustration of the principal eigenvector of citation correlation matrix $M'M$. (left) The result of increasing authority by spreading activation. Note that the authority of document x is not affected much by spreading activation. (right) Realignment of principal eigenvector by lifting of target document t. Authority of x greatly increases, because it is "just like t, only more so."

Let us say that we wish to align the principal eigenvector with document j. We begin by running HITS to convergence on the original link matrix. As described above, authority vector a converges to the principal eigenvector of $M'M$ such that iterative updates of

$$a^{+} = M'Ma,$$

yield $a^{+} = a$, where the $+$ superscript refers to the following time step. In particular a_j, the authority of document j, converges to a fixed point represented by

$$a_j^+ \quad = \quad \sum_i \left(\sum_k M_{ki} M_{kj} \right) a_j. \qquad (3)$$

Differentiating of Equation 3 with respect to link matrix entry M_{ki} gives

$$\frac{\partial a_j^+}{\partial M_{ki}} \quad = \quad M_{kj} a_i.$$

To perform gradient ascent, we simply add a fraction γ of the gradient to each element of the link matrix. After re-normalizing the links values, we re-run HITS and repeat. First, we apply HITS to compute authorities a from link matrix M, then we repeat the following steps:

1) Compute gradient $\forall_{k,i} \Delta M_{ki} = M_{kj} a_i$,
2) Update $M_{ki}^+ = M_{ki} + \gamma \cdot \Delta M_{ki} / \sum_i \Delta M_{ki}$,
3) Normalize weights, setting all $M_{ki} \geq 0$,
4) Re-compute HITS authorities a *using updated* M^+.

The normalization step is necessary since, by following the gradient, we may generate negative link weightings, implying that citation by a particular document actually *hurts* the authority of the cited document.[1]

For convenience, we refer to paper j as a "handle" that we "lift" to draw the desired documents away from the rest. Note that reversing the process is also straightforward – by subtracting the gradient, we decrease the authority of a document and move the principal eigenvector away from it.

The simplicity of the algorithm is appealing: the alignment of the principal eigenvector (and thus the principal topic and authorities) of the corpus is governed by the statistics of the citation matrix. As we perform gradient descent to align this eigenvector with a particular topic of interest, we implicitly encode this realignment by re-weighting the "importance" of citations in the matrix. Just as with gradient descent on the weights of a neural network, the algorithm achieves its objective without requiring the user to manipulate, or even be aware of the underlying weights.

When the method is applied to multiple handles at once (by adding the gradient terms of each handle) there is an important side effect: the elements that all handles have in common are accentuated, and those on which they differ tend to cancel. The result is a form of generalization – documents that rise in authority are those that are consistent with *all* lifted handles. Generalization, and the number of examples needed to achieve it, are well-studied (i.e. Blumer et al., 1989); future work will apply existing theoretical guarantees to our algorithm.

[1] This might be considered a feature by some, but complicates analysis of the system.

5. Experiments

Given the subjective nature of "authority," one of the greatest difficulties of introducing this sort of algorithm is evaluating it objectively. In this section, we describe two sets of experiments, both centered around generating customized authority lists of documents by altering the weights of the link matrix. In the first set of experiments we examine the result of lifting documents based on their *Cora* classifications; in the second set, we perform a user study on the quality of automatically-generated customized authority lists.

Our experiments focus on the "Machine Learning" portion of the *Cora* archive hierarchy. Under this topic, there are 4153 research papers categorized into six subtopics: "Case-Based," "Genetic Algorithms," "Neural Networks," "Probabilistic Methods," "Reinforcement Learning," and "Theory."

5.1 Evaluation of Lifting Based on Cora Classification

We can use the subtopic categorization of *Cora* as a form of objective test set, indicating relationship between two papers independent of their citation patterns. To evaluate our algorithm, we first consider the case of a researcher interested in only one particular subtopic of machine learning, say "Theory." Since theory papers are often under-cited in the experimental literature, none of the 512 theory papers in the archive may be considered "authoritative" with respect to the entire hierarchy, or even with respect to machine learning. Nonetheless, it is reasonable to expect that our theoretically-inclined researcher believes that theory papers are important, and given an arbitrary theory paper, would rank its authority with respect to machine learning higher than someone who was uninterested in theory. In the terminology of Bibliometrics, theory papers lie somewhere away from the top of the principal eigenvector for machine learning.

Working from this hypothesis, we could hope that by realigning the principal eigenvector to raise the authority of a small number of theory papers, we would see the authority of many other theory papers would rise.

If the goal was to classify papers in to subtopics, then there are other better ways to accomplish this. Instead, we are using subtopics as an objective indicator that we expect to be correlated with a user's interests.

Beginning with a root set of all documents in the "Machine Learning" topic, we ran 10 iterations of HITS to establish initial authority rankings. We selected a small number of papers (between 10 and 20) from one of six subtopics, and designated them as authorities. We then alternated one step of lifting (step size $\gamma = 0.01$) with 3 further iterations of HITS, and observed the ranks of the lifted set and of all papers in the subtopic from which it was drawn.

We studied two methods for drawing a lifted set from its subtopic. One set of experiments drew the lifted set at random from the selected subtopic. The other selected those papers in the subtopic that were most highly ranked by the initial application of HITS, attempting to mimic a user's selection of authoritative papers in their preferred area.

As expected, lifting a small number of papers from any of the topics produced an amplified improvement in the ranks of other papers in the topic (Figure 2). There were, however, interesting and unanticipated variations in the results.

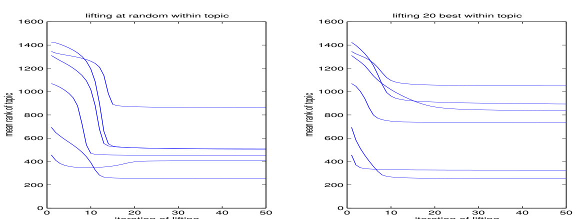

Figure 2. Mean rank of papers in different subtopics as members are lifted. With the exception of "Genetic Algorithms," the effect of lifting a small number of papers is amplified into a large improvement in rank for the topic.

We found little qualitative difference between the rank modification curves as a function of the number of papers lifted. Lifting more papers made the effect more dramatic, but had no qualitative effect on the shape of the curve.

The selection criterion for the lifted papers (n-best vs. n-random) had a significant effect on the shape of the rank modification curves. When the n-highest ranked papers in a topic were lifted, a large subset of the other papers in the topic also increased in rank. When the papers to be lifted were chosen at random, almost all of the papers in the topic increased rank, but in many cases did not increase as much as in n-best experiment.

Independent of the number of papers lifted and the selection criterion for choosing them, we found surprising variety of shapes of the rank modification curves, which, we suspect, reflects the different citation patterns of the corresponding research communities.

Some fields, such as Reinforcement Learning, appear cohesive and tightly linked – lifting any reinforcement papers brings them all to the forefront (Figure 3). Other communities, such as Case-Based Learning, seem to be split – lifting the most "authoritative" papers has a large effect on some, but leaves an equal number almost untouched (Figure 4). Most unusual was Genetic Algorithms (Figure 5), in which lifting some papers significantly *decreased* the rank of many others in the same area. A full understanding of what is implied by these differently shaped curves is a subject of future work.

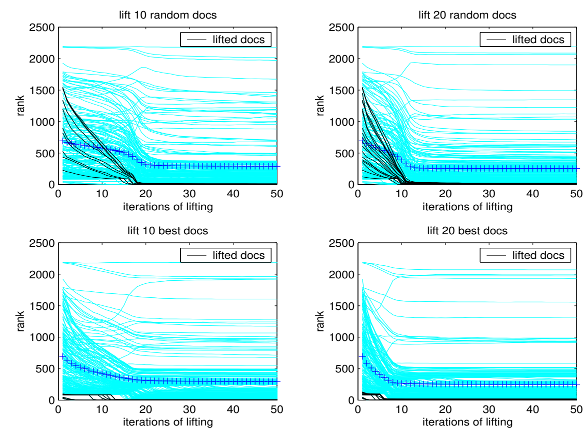

Figure 3. Rank of RL documents as a subset is lifted. General movement of most papers in topic indicates a tightly-knit, cohesive research area. Crosshatched line indicates mean topic rank.

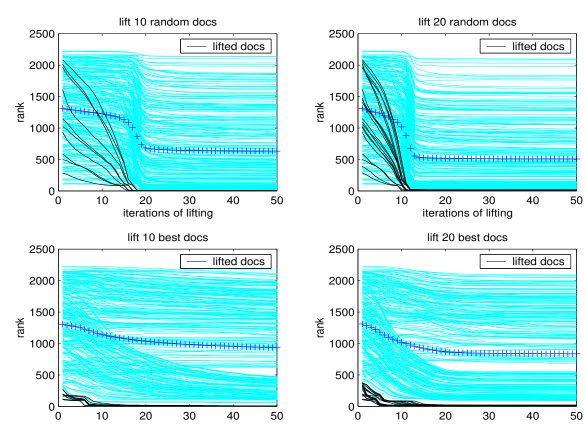

Figure 4. Rank of case-based learning documents as a subset is lifted. Stratified movement of papers suggests multiple communities within the area. Crosshatched line indicates mean topic rank.

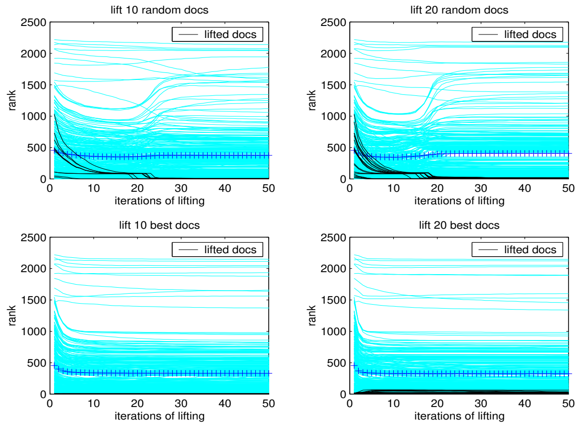

Figure 5. Rank of genetic algorithm documents as a subset is lifted. Crosshatched line indicates mean topic rank.

5.2 Evaluation of Lifting based on Human Subjects

The goal of the second set of experiments was to automatically create customized authority lists. We created a customized authority list for a person by finding the ten most-cited papers they have authored, "lifting" them, and running our modified HITS algorithm.

The experiment was performed by making customized authority lists for each subject, pairing that subject with someone from their group, and merging their authority lists so as to create ten pairs of references: the top authority from each subject, the second authority from each subject, and so on. Each subject was sent the merged list containing their customized authorities, and asked to indicated which reference of each pair was "most authoritative and relevant to their own personal area of research." When they selected the reference from the authority list we customized for them, this was counted as correct; when they selected the other, this was counted as incorrect; when they indicated no preference, this was not counted. From these counts, percentage accuracy was calculated.

We contacted fourteen prominent researchers in Machine Learning by email and asked them to take part in our study. Of these fourteen, two did not respond, and two responded saying that the lists of pairs did not make sense to them. The remaining ten fall into two categories: subjects A, B, C, D, E, and F all do research in Machine Learning, but not the same sub-area of Machine Learning; subjects W, X, Y, and Z all do research in the Reinforcement Learning (RL) sub-area of Machine Learning. The results are as follows: A 84%, B 66%, C 89%, D 73%, E 89%, F 57%, W 88%, X 19%, Y 50%, Z 31%.

The overall accuracy was 64.6% (statistically significant with 95% confidence, using a one-sided t-test and $\alpha = 0.05$). The average accuracy among the non-RL subjects is even higher (76.3%, significant with grater than 99% confidence), indicating that we can indeed create authority lists successfully customized to a particular person's notion of authority. Interestingly, examination of the lists shows papers from multiple sub-areas of Machine Learning when that person does research in multiple sub-areas. Furthermore, as described earlier, the lists show more than simply spreading activation. The lists are not simply papers near the subjects' publications, but contain papers at the convergence of the authors' interests.

Among the RL subjects average accuracy was much lower: 47%. It is reasonable that it would be more difficult to separate people with such tightly-related interests. What is most interesting however, is the two subjects W (88%) and X (19%) who were paired together. Subject W is well established in the field and has been publishing in RL for more than ten years; subject X is relatively new to the field and has fewer publications. From the results, we see that subject X prefers the authorities found by W to his own authorities. We believe this is exactly an example of the phenomena pictured in Figure 1. That is, subject W is like X, only more so.

Examination of the link matrix reveals that X is heavily cited *outside* of reinforcement learning, and that *Cora* contains a number of his non-RL papers. This "authority" outside of RL actually competes with the authority of his RL papers, and may also explain the results we have observed. In cases when a person specializes in multiple areas, it is not clear that they will will always be interested in the mixture of these areas.

These observations highlight another problem for lifting (and all algorithms based on link structures): the lack of consideration for chronology. As a researcher's interests change throughout their career, a list based only on their "most popular" papers may not reflect their current interests. One possible fix is to use only customized handles selected by the user to reflect their current interests. For example, we could ask X to select papers that he *currently* considers authorities, and lift those as handles.

One further chronological factor affects the performance of HITS. Older papers will have more references to them; while newer papers will have a better view of the "intellectual space," making it easier to create good "hub" documents. Other than measuring link rate (links per unit time) instead of raw counts, it is not clear how such an effect may be managed.

6. Related Work

6.1 Bibliometrics

Bibliometrics is the study of literatures as they are reflected by the citations (White & McCain, 1989). Author co-citation analysis (ACA) exploits the link structure of journal articles to identify groups such as subject areas, research specialties, schools of thought, shared intellectual styles, or temporal geographic ties (Bayer et al., 1990; McCain, 1990). To perform ACA, a co-citation matrix is first formed. Two authors have a co-citation if a third author makes reference to both of them in the same paper. This corresponds to $M'M$ in HITS. (Similarly, MM' corresponds to a matrix containing "bibliographic coupling" counts, where two papers are bibliographically coupled if they citations they make have papers in common.) It is often preferable to analyze the correlation matrix of co-citation data. This matrix contain similarity measure of the citation profile for each pair of authors. The profile provides more information about co-citation relations of authors than the raw frequency matrix, and it attenuates the scale difference between a highly cited author and others who are less popular but have similar profile. Any of three popular techniques can be applied to the matrix:

Papers can be clustered, using any clustering technique, to find groupings of the papers that are indicative of intellectual organization within the field of study. Entries of the correlation matrix are used as a similarity measure. The nesting of clusters can yield useful information about relation of fields and their sub-fields.

Multidimensional scaling methods can give the analyst a visualization of the intellectual space by providing information rich two dimensional display of the data. Again, the entries in the matrix are used as distance measure. Authors writing on the same subject will tend to be co-cited most often, and their position on a plot will show them to be at center of a group. Those who overlap less with those at the center of a group will tend toward the periphery.

Factor analysis of the matrix can identify groups and relative importance of each author within every group. Principal components analysis is often used to extract factors. An author's loading on a factor measures his contribution to the group (the authority score in HITS is equivalent to "loading" of an author on the principle eigenvector). Since authors may have significant loading on more than one factor, this method also reveals the "breadth" of an author's writing.

Traditionally, ACA is done "by hand," that is, journal citation data are used to create the an appropriate matrix, and a statistical package such as SPSS-X or SAS is used to apply the above methods. The results are then interpreted by a Bibliometrics analyst, who matches the groups of authors provided by each method with a subject areas or specializations. This type of analysis is an effective way to analyze published material, and is still performed today, (e.g., Raghupathi & Nerur, 1999).

6.2 Learning Rankings

The general problem of learning to rank objects according to user preferences has received a great deal of attention (e.g., see Cohen et al., 1999). Our lifting algorithm is constrained to generate rankings from a linear transformation of the co-citation matrix. As such, it is only able to generate a small subset of the possible rankings that a user might want. On the other hand, it does not suffer from the computational complexity of the general ranking problem. It would be interesting to apply lifting to a more general model of document authority to determine whether the widely varying performance of the lifting algorithm is due to the restricted model or some other factor.

6.3 Relevance Feedback

Our method of interacting with the user closely resembles relevance feedback (van Rijsbergen, 1979), a technique used for information retrieval. In relevance feedback, the user poses an initial query to the retrieval engine, which returns a set of candidate documents. The user may then score these documents, by voting for "more like this" or "less like this." The retrieval engine then expands the the initial query, looking for documents that more closely resemble the high-ranking documents from the initial set, and rejecting those that resemble the low ranking ones. The process may be repeated, allowing the user to score documents on each iteration, until the query has been refined to return just the documents that the user wants.

Our lifting method may be said to accomplish the same thing for Bibliometrics as relevance feedback accomplishes for information retrieval: it allows system to iteratively refines its model of "what the user wants." Where relevance feedback modifies the weighting of terms in a document to influence its relevance, we modify the weighting of citations in the link matrix to influence its authority. Given the recent confluence of information retrieval and Bibliometrics on the web, we expect that these two techniques would be especially powerful when used in parallel.

7. Future Work

An obvious application of lifting is the dynamic generation of customized web page search results. Given precomputed HITS scores, we can efficiently lift a small set of user-specified documents and generate customized scores in real time. Moreover, we expect great enhancement of lifting results due to the web's *much richer* link structure, and more numerous authority pages with finer distinctions.

More generally, lifting can be applied to other "spectral analysis" based methods (Chakrabarti et al., 1998). Given any preference-based matrix, we can incorporate user feedback by lifting the chosen handles. Again, in such an information dense environment, we expect lifting to provide even more dramatic results.

As indicated, it would be helpful to have some form of performance guarantees on how many documents must to be lifted (or lowered) to align the eigenvectors of a particular link matrix. We expect that such guarantees may be derived by computing the Vapnik-Chervonenkis dimension of the system (Blumer et al., 1989), and establishing a numerical loss function to which the bounds could be applied.

8. Summary

Much information is provided in the link structures of corpora such as the web, or archives of scientific publications. In this paper, we have introduced a method of combining information gleaned from these structures with feedback from a human user to learn the internal model of authoritativeness of that user. By "lifting" a few feedback examples, we can create customized authority list. The method is both time and space efficient, and only relies information stored in the links.

Our experiments have shown that on a corpus of computer science research papers, using only a small amount of user feedback, lifting is able to significantly improve HITS ranking so that it is in better agreement with both non-link based machine classification and human subjects' preferences.

Acknowledgments

Many thanks to the following people for participating in our user study: Craig Boutilier, Avrim Blum, Justin Boyan, Tom Dietterich, Pedro Domingos, Yoav Freund, Nir Friedman, Thomas Hofmann, Pat Langley, Michael Littman, Andrew Moore, Andrew Ng, Justinian Rosca, Yoram Singer, Rich Sutton and Sebastian Thrun; and also many thanks to those who wish to remain anonymous.

References

Bayer, A. E., Smart, J. C., & McLaughlin, G. (1990). Mapping intellectual structure of a scientific subfield through author cocitations. *Journal of the American Society for Information Science*, 433–443.

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

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of Association of Computing Machinery*, *36*, 929–965.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh World-Wide Web Conference.*

Chakrabarti, S., Dom, B. E., Gibson, D., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1998). Spectral filtering for resource discovery. *ACM-SIGIR Workshop on Hypertext IR for the Web.*

Cohen, W., Schapire, R., & Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, *10*, 243–270.

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

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

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41,6*, 430–432.

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

McCallum, A., Nigam, K., & Ungar, L. (in press). Efficient clustering of high-dimensional data sets with application to reference matching. *Sixth ACM-SIGKDD Conference on Knowledge Discovery and Data Mining.*

Raghupathi, W., & Nerur, S. P. (1999). Research themes and trends in artificial intelligence: An author cocitation analysis. *Intelligence-New visions of AI in Practice*, 18–28.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–289.

van Rijsbergen, C. J. (1979). *Information retrieval.* London: Butterworths.

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