Author Identification from Citations

Joseph K. Bradley, Patrick Gage Kelley, and Aaron Roth
Department of Machine Learning
Institute for Software Research
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213
(jkbradle, pkelley, aroth)@cs.cmu.edu

Abstract

Machine Learning techniques can be applied to citation data from a network of papers to predict the author of a paper that is currently outside of the network. Using a series of models we have found that we can increase the accuracy from past experiments with citation data, by considering the citations as a network. This allows us to predict with confidence the author of a blind paper.

1 Introduction

Blind or masked review of papers for both conferences and journals is a practice that is used to guarantee equality among all authors and institutes who submit papers.

However, this method of review long been questioned as actually creating anonymity for authors, for papers can often be recognized by topic area, notation, methods, and knowledge by a reviewer from a similar field.

Our main question here is how easy it truly is to determine the author of such a blind paper. Using only the references and citation network that the paper fits into, which is publicly available, we explore the ability of a machine learning system to automatically discover the author of this paper.

2 Problem Definition

The goal of this project is to create a system which, when given an unpublished paper with the names of the authors and affiliations removed, will identify with some confidence the author(s) who wrote the paper. We base these predictions on the frequency of cited authors and papers and the strucures formed in the citation graph based on specific sub-topic.

A system which can automatically determine the author of an anonymous paper with high accuracy would clearly show that blind review is in fact not blind and needs to be replaced with a better system.

3 Method

3.1 Intuition

The problem of author identification for academic papers has received limited attention in the past. To our knowledge, Hill and Provost [1] made the only significant contribution to solving this problem. They proposed this task and made preliminary attempts to implement a solution as their entry into the KDD Cup 2003 competition. They worked with the KDD Cup citation dataset: a large collection of physics papers from arXiv. They did not use any textual features, instead considering only the citation information for each paper.

Hill and Provost consider only two simple methods. One technique they consider is to represent each paper as a vector of citations over the space of the entire dataset, and representing each author as the sum of the vectors of each of his papers. They experiment with several vector representations: binary citation values, citation counts, and citation counts weighted by a decay term corresponding to the age of the paper. They simply classify new papers as having been written by the author to which the paper vector matches most closely, using cosine similarity to measure vector distance. This technique achieves 26% accuracy.

The other method they try is to predict naively that each paper is written by the author that is most cited in that paper. This approach suffers from the fact that it can never make a correct prediction unless the author has cited himself at least two times, but in fact, it achieves 37% accuracy.

Since Hill and Provost achieve moderately good results without using any machine learning, except for the memory-based vector similarity approach, it should be possible to achieve a significant performance increase using more sophisticated techniques.

In the following sections, we describe our dataset and propose several machine learning methods for addressing this problem.

3.2 CiteSeer Data

Full citation data is available from CiteSeer as mentioned. After downloading this data, we processed it and found that the main database has 716,772 papers, 337,118 of which have citations, with an average count of 5.24 citations per paper. By requiring that papers have at least 5 citations, we built a working set of 137,485 paper with an average citation count of 9.67.

The two main datasets discussed in this paper are portions of the whole CiteSeer data. The set used in most of our graphical representations of the data is an expansion of the citations based off of a single node: CiteSeer paper number 3745, Trace-Driven Memory Simulation: A Survey, by Richard A. Uhlig. This set then expands outward based on the citations of this paper adding an additional 27 papers in the first iteration, with a total of 109, 287, and 567 papers by the second, third, and fourth iterations. The second data set is an extraction of papers in the CiteSeer database; it was built by selecting all papers from NIPS and all papers connected to NIPS via citations. This results in 42,978 papers and is used in the algorithms below.

3.3 Latent Semantic Analysis

Our first approach to learning mappings from citations to authors uses Deerwester et al.'s Latent Semantic Analysis (LSA) method, which was originally developed for modeling documents for information retrieval [3]. LSA gives an efficient representation of a document collection which may then be used to find documents similar to some new document or to find documents relevant to a set of search terms. Previous methods[1] often used sim-

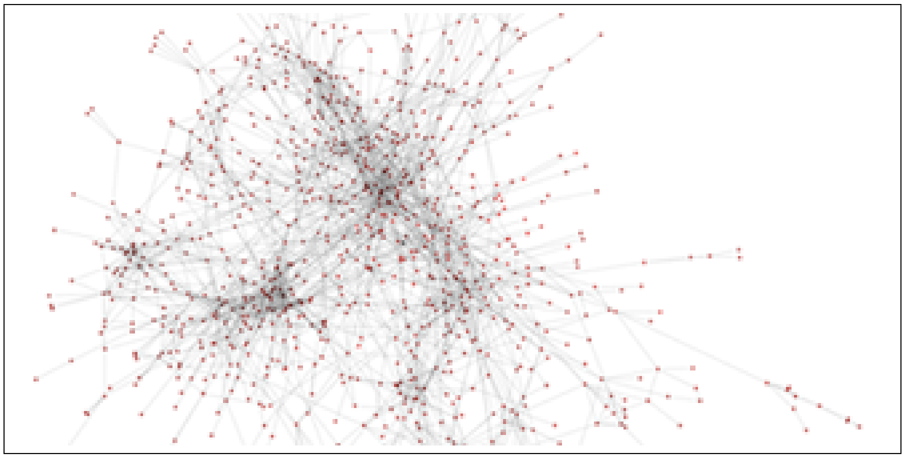

Figure 1: Graphical Data Representation of Citation Expansion

ple term-based comparisons which were inefficient and failed to account for phenomena such as synonymy (distinct terms with the same meaning) and polysemy (one term with multiple meanings); LSA partially addresses these issues.

Deerwester et al. argue that, in addition to allowing future queries to be computed more efficiently, the rank reduction accounts for synonymous terms in the documents. Synonymous terms tend to be nearly linearly dependent in the original term-document matrix M, so performing SVD and rank reduction maps synonymous terms to the same point in the reduced topic space. They note that LSA is less successful at dealing with polysemy than with synonymy.

To see how we apply LSA to our problem of predicting authors from citation data, first suppose that each author has written only one paper in the training set. The natural way to apply LSA then is to interpret authors as terms and papers as documents. Since each author has written only one paper in the author/paper matrix, choosing the original paper receiving the highest similarity gives a set of possible authors for a new paper.

To eliminate the assumption that each author has written only one paper in the training data, we interpret documents as possible authors instead of papers. In other words, a document in LSA becomes a possible author a in our method, and a is represented by a vector of weights over other authors whom a has cited.

This method does lose some information by combining all of an author's papers into one weight vector for that author; this is equivalent to ignoring polysemy, or the fact that an author may write about several different topics. However, this method still takes advantage of LSA's ability to deal with synonymy, where multiple authors may write about similar topics.

3.4 Latent Dirichlet Allocation

We next addressed our problem of author prediction using the language of graphical models. While LSA has no ready probabilistic interpretation, we can make probabilistically meaningful predictions using graphical models. We chose to apply the Latent Dirichlet Allocation (LDA) model of Blei et al. [5] in a manner similar to how we applied LSA. Below, we present our formulation of the problem, which is identical to LDA except that topics are replaced with authors and words replaced with cited authors, and we derive inference algorithms. Though we do not present results from the model, we predict that it will perform better than LSA since Blei et al. were able to show that LDA handles issues such as synonymy and polysemy more effectively than LSA.

We describe a paper i as a distribution θ_i over K possible authors who wrote the paper and a bag of (possibly repeated) cited authors $\mathbf{w}_{i,n}$. Each cited author is represented as a C-vector with a 1 in the cited author's position and 0s elsewhere. In this model, a paper is generated as follows: The distribution over a paper's authors θ_i is drawn from a Dirichlet distribution with parameter α. For each of the N_i cited authors, an author $\mathbf{z}_{i,n}$ of the paper is chosen from a $Multinomial(1, \theta_i)$ distribution, where $\mathbf{z}_{i,n}$ is represented as a K-vector in the same way as $\mathbf{w}_{i,n}$; this author then chooses an author $\mathbf{w}_{i,n}$ to cite according to a $Multinomial(1, \beta_{z_{i,n}})$ distribution. We can interpret $\beta_{z,c}$ as the probability that author z cites author c.

We first consider the problem of inferring the model parameters α and β. Given a set of training papers, we can observe θ_i and $\mathbf{w}_{i,n}$; this is the main difference between this problem and Blei et al.'s application of LDA: in LDA, θ_i is never observed. As Blei et al. show for LDA, exact inference is intractable for this model, so we derive an Expectation-Maximization (EM) algorithm for parameter estimation below. We omit much of the proof due to space constraints.

Dropping the paper subscript i, the joint probability of a paper is

$$\begin{aligned} p\left(\theta, \mathbf{w} | \alpha, \beta\right) &= \sum_{\mathbf{z}} p\left(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta\right) \\ &= \sum_{\mathbf{z}} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right) \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \theta_k^{z_{n,k}}\right) \left(\prod_{c=1}^{C} \beta_{z_{n,c}}^{w_{n,c}}\right) \end{aligned}$$

Since θ is observed, we can estimate α separately from β. To estimate α, we can take the derivative of the log likelihood and show that the Hessian is of the form required to use the Newton-Raphson method described by Blei et al. to maximize the likelihood with respect to α. To estimate β, we can use a variational inference algorithm. We approximate the distribution $p(\mathbf{z}|\theta, \mathbf{w}, \alpha, \beta)$ with the multinomial distribution $q(\mathbf{z}|\phi)$. To maximize the likelihood $p(\mathbf{w}, \theta|\alpha, \beta)$ of a paper, we lower-bound this likelihood using Jensen's inequality:

$$\begin{aligned} \log p(\mathbf{w}, \theta | \alpha, \beta) &= \log \sum_{\mathbf{z}} \frac{p(\mathbf{w}, \theta|\alpha, \beta) q(\mathbf{z}|\phi)}{q(\mathbf{z}|\phi)} \\ &\geq \mathrm{E}_q[\log p(\mathbf{w}, \theta|\alpha, \beta)] - \mathrm{E}_q[q(\mathbf{z}|\phi)] \end{aligned}$$

Letting the above lower bound equal $L(\phi; \beta)$, we can show that

$$\begin{aligned} L(\phi; \beta) &= \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} \log \theta_k\right) + \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{c=1}^{C} \phi_{n,k} w_{n,c} \log \beta_{k,c}\right) \\ &\quad - \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} \log \phi_{n,k}\right) \end{aligned}$$

Adding Lagrange multipliers to ensure the constraints that $\sum_{k=1}^{K} \phi_{n,k} = 1, \forall n$, and that $\sum_{c=1}^{C} \beta_{k,c} = 1, \forall k$, we can take the derivatives with respect to $\phi_{n,k}$ and $\beta_{k,c}$ and set them equal to zero. Solving the resulting set of equations gives update rules for an iterative algorithm which converges to a locally optimal estimate of β. We let $w_n^{(i)} = v$ s.t. $w_{n,v}^{(i)} = 1$.

$$\begin{aligned}\phi_{n,k}^{(i)} &\propto \theta_k^{(i)}\beta_{k,w_n^{(i)}} \quad \forall\, \text{papers}\, i \\ \beta_{k,v} &= \sum_{i=1}^{M}\sum_{n=1}^{N_i}\phi_{n,k}^{(i)}w_{n,v}^{(i)}\end{aligned}$$

Once we have estimated the model parameters, we may perform inference on a yet unseen paper by estimating the now-hidden variable θ from the observed cited authors. This inference problem is identical to the one addressed by Blei et al. and may be solved using the same algorithm.

3.5 Characteristic Vector Classifier

We use the intuition that authors write many papers with similar citation "signatures". We consider variations on two types of signatures for each paper. The first is a vector with one entry for every paper in our dataset, and the second is a vector with one entry for each author. For both types of vectors, we consider signatures of varying depths, d. For $d = 1$, our signatures are simply a record of the references in a given paper. For example, entry $x_i = 1$ if the paper sites paper i in the case of paper vectors, and $x_i = k$ where k is the number of times author k is cited in the case of author vectors. For $d = 1$, this is similar to the vectors used by Provost and Hill. (We note however that they compare individual papers to author-averages, whereas we compare papers directly to other papers). For depth $d > 1$, we perform a depth d traversal of the citation graph starting at our initial paper, and taking each possible directed walk of length d, maintaining counts of authors and papers to attain our characteristic vector.

When classifying a paper x, we proceed as follows: Given the reference set for x, we compute its characteristic vector, and then compare it using cosine similarity to each other paper y:

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

We then sort the papers by similarity, and identify each paper with its author set. We therefore have an ordering of authors in order of the similarity of their papers to the target paper x, but in this ordering, each author may appear multiple times (since they may have multiple similar papers). We give each author a single ranking as follows. If Y is the set of indices for which author a appears in our list (1 being the best), we compute the score for author a:

$$\texttt{Score}(a) = \sum_{i \in Y} 500000 \cdot (\frac{1}{1.09})^i$$

We then order each author according to their aggregate score to produce our predictions.

3.6 Random Walk Classifier

We use the intuition that authors write many papers on the same topic, with similar connections in the citation graph to motivate this approach. Given a paper P by an unknown author with a set of references, we construct a probability distribution over the set of authors by the following method:

1. Repeat the following $10,000$ times:

2. Pick a random paper r that P cites.
3. Pick a random paper s that cites r (excluding P).
4. Let $r = s$. With probability $2/3$, HALT and add probability mass to each author of r. with probability $1/3$, go to step 3.

In this way, we achieve a probability distribution over authors defined to be the probability that a random walk of expected length 1.5 lands on these authors starting at one of the papers cited by our target paper p. We then rank the authors by their probability, and report the top 20.

4 Experiments

4.1 Experimental Questions

Our experiments hereafter contain a series of questions, all stemming from the main goal of the project.

Given an unidentified paper, as it would be submitted for a blind review, is it possible to construct a machine learning solution that can correctly identify the author(s) of such a paper?

However, other questions have derived from our initial work and conversations about the initial structure:

Given a set of citation data, do the citations form subgroups or have some sort of structure?

Given a set of citations, what additional citations would be beneficial or what authors? work should be read?

This final question is a direction that shows our work in a positive light, as it would provide researchers who have located a few specific relevant papers with other related directions that they may not have otherwise found.

4.2 Details

We implemented LSA, the Characteristic Vector Comparison, and Random Walk methods and then tested them on sets of papers taken from the CiteSeer dataset. For LSA, we used a subset of the NIPS-based dataset described above, built by taking NIPS papers and all papers cited in a NIPS paper; for the other methods, we used the entire NIPS-based dataset. We only consider authors who wrote NIPS or ICML papers or were cited in the papers. For LSA, to divide the papers into training and test sets, we randomly permuted the papers and then selected the first 2/3 of them for training, reserving the rest for testing. For the other methods, we used leave-one-out cross-validation.

Each of these methods resulted in accuracies that are much higher than the past work had found. The results are detailed in the table below. The percentages give the percent of true authors which are predicted within the first N predicted authors, where N is given in the following column. Therefore, answering our first research question, we have shown that it is possible to create multiple machine learning systems which can predict a set that includes the author of a given paper with over 50% accuracy. Further work on refining these algorithms or adding additional features could result in even higher accuracy (lower recall - higher precision results).

With our second question it is shown in Figure 1, representing a portion of that data that given a set of citation data, we have observed smaller cycles in the paths as well as larger formations of divisions between fields. This is apparent in the visual representations by darker areas, where these graphs have been computed by using edges as springs pulling connected nodes together. This is also shown by falsely high scoring results which are frequently researchers working on very similar topics.

Using a given set of citations we can now create a reading list of papers that a researcher should look to for further work, as well as a list of authors who have made substantial contributions to an area. This method finds the most likely papers that are citing a set of input papers, which would be a reasonable list of works which are relevant to the current research.

4.3 Results Table

Table 1: Results Returned Comparison

Method	Details	Accuracy	# Predictions
Hill & Provost	Past Work	25%-45%	
LSA	150 Latent Topics	64.96%	50
Characteristic	Depth 1	62.24%	32
Comparison	Depth 2	68.91%	52
(Paper)	Depth 3	68.85%	55
Characteristic	Depth 1	67.27%	58
Comparison	Depth 2	70.48%	68
(Author)	Depth 3	69.98%	69
Random	10000 runs & 2/3 stopping probability	56.27%	37
Walk	10000 runs & 1/3 stopping probability	56.52%	42

5 Future Work

To improve on our work, many of the given algorithms could be refined, and the features from each could be combined to provide an overall response. Using the LDA model described above would also likely yield substantial improvement, for the LDA model has more firm statistical foundations than more heuristic methods such as LSA. Our prediction methods could then be posted as a web-page that would allow referees to input the references from a blind paper they are reviewing, and then be given a list of probable authors.

6 Conclusions

The main portion of this project was creating multiple machine learning systems which, given a set of papers in a field, learns a mapping from anonymous papers to possible authors, ranked with confidences. From our initial results, we are confident that this system can be expanded into a tool to complete this task for true unknown papers in a public space.

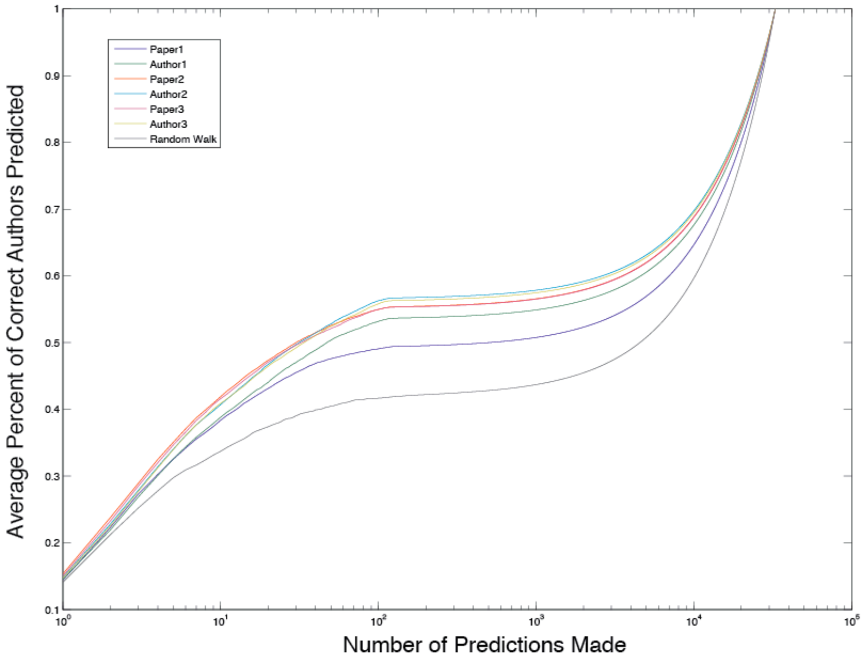

Figure 2: Recall Curves Over Multiple Authors

Our initial results have also shown that our methods may be used to provide un-included citations that are likely valuable. For example, for a paper that cites a number of papers in the graph, we can suggest a number of papers that are relevant to the work that has been done, by those authors, or in relation to that work.

References

[1] *The Myth of the Double-Blind Review?: Author Identification Using Only Citations. Foster Provost and Shawndra Hill - http://portal.acm.org/citation.cfm?id=981001*

[2] Properties of Academic Paper References. Sunghun Kim, E. James Whitehead, Jr. http://www.cse.ucsc.edu/ ejw/papers/kim_ht04a.pdf

[3] Indexing by Latent Semantic Analysis. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, R. Harshman. http://citeseer.ist.psu.edu/deerwester90indexing

[4] Information Bottleneck Method. Naftali Tishby. Fernando C. Pereira. William Bialek. http://www.cis.upenn.edu/ pereira/papers/allerton.pdf

[5] ”Latent Dirichlet allocation.” Blei, D., A. Ng, and M. Jordan. Journal of Machine Learning Research, 3:993-1022, 2003.