Mining a Digital Library for Influential Authors

David Mimno, Andrew McCallum
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA
{mimno,mccallum}@cs.umass.edu

ABSTRACT

When browsing a digital library of research papers, it is natural to ask which authors are most influential in a particular topic. We present a probabilistic model that ranks authors based on their influence in particular areas of scientific research. This model combines several sources of information: citation information between documents as represented by PageRank scores, authorship data gathered through automatic information extraction, and the words in paper abstracts. We compare the performance of a topic model versus a smoothed language model by assessing the number of major award winners in the resulting ranked list of researchers.

Categories and Subject Desriptors: H.3.7 Information Systems : Digital Libraries **General Terms:** Algorithms.

Keywords: Expert Retrieval.

1. INTRODUCTION

Measuring the influence of researchers is an important task. The current trend towards open-access electronic publishing in academic disciplines has made it increasingly possible to derive various indicators of impact from research literature. It is necessary, however, to put such metrics in context. An academic digital library collection may contain many subdisciplines, each of which has its own influential researchers. In this paper, we demonstrate one method for finding domain experts using the Rexa Digital Library [3].

In order to find experts for a given topical query, we combine information from several sources, all derived from scientific papers available on the web, including titles, abstracts and author names extracted from PDF documents. Authors are coreferenced using Machine Learning methods. Previous studies of author influence in digital library collections have been hampered by ambiguities in authorship (for example, Newman [4] groups authors by first initial and last name). Finally, the link structure of the collection is identified by extracting and disambiguating the references from papers.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.

We compare three statistical models for the association of a text query with the words in titles and abstracts of papers by each author. We evaluate each model by comparing the resulting ranked lists of authors with recipients of major awards in the specified topical area.

2. METHODS

We measure the influence of individual documents using the PageRank algorithm. Chen et al. [2] demonstrate the use of PageRank on research literature, using references in place of hyperlinks. PageRank can be thought of as modeling a researcher who moves from paper to paper in the document collection. At each paper, the researcher either follows a randomly chosen reference from the current paper or, with probability δ, chooses a random paper from the collection. The PageRank of a given paper can be interpreted as the probability that the researcher will be reading that paper at any given moment. Since the PageRank is a probability distribution over all documents in the collection, we use it as the probability of a given document, $\Pr(d)$. We follow Chen et al. [2] in their use of $\delta = 0.5$, representing the assumption that readers of scientific papers are more likely to jump to a random new document than web surfers.

For the probability of authors given documents we use a uniform distribution, dividing the weight of a document evenly between its authors. The probability of an author given a document is $\Pr(a|d) = 1/(|A_d|)$, where A_d is the set of authors in paper d.

Using these elements we can construct a distribution over authors for a particular query,

$$\Pr(a|q) \quad \propto \quad \sum_d \frac{1}{|A_d|} I_{\{a \in A_d\}} \Pr(d) \Pr(q|d) \qquad (1)$$

where $I_{\{a \in A_d\}}$ indicates whether a is listed as an author for a given paper. For the component of the model that depends on the words in documents, $\Pr(q|d)$, we compare three statistical models. The first is based on a language model with Dirichlet smoothing. The second and third are based on a statistical topic model, using a single topic and a weighted sum of topics, respectively. Recent work by Wei and Croft [5] shows that interpolations of language models and topic models improve performance in information retrieval. In this work, we examine each of these components separately.

For the language model with Dirichlet smoothing, the probability of a query given a document is

$$\Pr(q|d) \quad = \quad \prod_{w \in q} \frac{N_d}{N_d + \mu} \frac{N_d^w}{N_d} + \frac{\mu}{N_d + \mu} \frac{N^w}{N} \qquad (2)$$

where $\mu = 100$, N_d is the number of words in document d, N_d^w is the number of times word w appears in document d, N^w is the number of times word w appears in the corpus, and N is the total number of tokens in the corpus. This smoothing allows documents that do not contain all query words to have non-zero probability and reduces the effect of very short documents that contain only query words.

For the topic model we use Latent Dirichlet Allocation (LDA) [1]. LDA models documents as mixtures of "topics", which are probability distributions over the vocabulary of the corpus. Topic models are useful in handling synonymy (multiple words with similar meanings) and polysemy (words with multiple meanings), because they assign words to topics based on the context of the document. In this application, the topic model can be considered a sort of query expansion: documents that contain none of the query words may still contain words that commonly occur in the same contexts as the query words. A trained topic model produces an estimate of the probability of a word given a topic, $\Pr(w|t)$, and the probability of a topic given a document, $\Pr(t|d)$. In the second model we select a single topic t that matches the query and substitute $\Pr(t|d)$ for $\Pr(q|d)$ in Equation 1. In the third model we represent $\Pr(q|d)$ as a weighted sum over all topics:

$$\Pr(q|d) \quad = \quad \prod_{w \in q} \sum_{t} \Pr(w|t) \Pr(t|d). \qquad (3)$$

We make the assumption that authors and words are conditionally independent given a particular document. In other words, if you specify a paper, knowing who wrote that paper tells you nothing about the words in the paper that you did not already know.

3. RESULTS

Influential researchers for the query "information retrieval" from the three models are shown in Table 1. The topic model for the second and third models consists of 400 topics trained on the Rexa corpus. For the second model, we manually select a topic indicated by "information, document, documents, retrieval, structured, ir, relevant, collections."

In order to evaluate the results, we highlight winners of three major awards. First, the Gerard Salton award, for "significant, sustained and continuing contributions to research in information retrieval" (van Rijsbergen, Croft, Robertson, Saracevic, Cooper, Sparck Jones, Salton).[1] Second, the Tony Kent Strix award, for "an outstanding contribution to the field of information retrieval" (van Rijsbergen, Harman, Robertson).[2] Third, the ASIS&T Award of Merit, for "a noteworthy contribution to the field of information science" (Belkin, Sparck Jones, Saracevic, Salton).[3] This method of identifying experts is crude: non-recipients of major awards are not necessarily less influential. We can, however, reasonably assume that those who have received an award should be considered experts.

Of the three models tested, the weighted topics model identifies nine award winners, including seven of the eight Salton award winners in its top thirty ranked authors. The language model identifies six award winners and the single topic model identifies five. The single topic model also ranks several authors who are only peripherally connected to information retrieval, such as Serge Abiteboul (databases) and Leslie Lamport (distributed computing).

The weighted topic model approach to expert finding appears to be better able to generalize beyond the specific query words, while retaining a focus on areas relevant to the query. We believe that such models are a promising direction in expert finding, and a good example of the usefulness of structured digital library collections.

Table 1: A comparison of authors ranked by each model for the query "information retrieval". The topic chosen for the second model is "information, document, documents, retrieval, structured, ir, relevant, collections." Recipients of the Salton, Strix and ASIS&T awards are marked in boldface.

Language Model (6 awards)	Single Topic (5 awards)	Weighted Topics (9 awards)
W Bruce Croft	**W Bruce Croft**	**W Bruce Croft**
Norbert Fuhr	**Keith van Rijsbergen**	**Keith van Rijsbergen**
Nicholas J Belkin	Norbert Fuhr	Chris Buckley
Douglas W Oard	Marti A Hearst	Norbert Fuhr
Fabio Crestani	**Nicholas J Belkin**	Ellen M Voorhees
Ellen M Voorhees	James P Callan	**Donna K Harman**
Keith van Rijsbergen	Craig A Knoblock	**Karen Sparck Jones**
Mounia Lalmas	Chris Buckley	**Gerard Salton**
Alexander Hauptmann	**Karen Sparck Jones**	Joseph John Rocchio
Chris Buckley	**Gerard Salton**	**Stephen E Robertson**
Fabrizio Sebastiani	Serge Abiteboul	David D Lewis
James P Callan	Douglas W Oard	**Nicholas J Belkin**
David D Lewis	Stuart K Card	Alan F Smeaton
Mark Sanderson	Frank Z Smadja	Marti A Hearst
Sandor Dominich	Howard R Turtle	Douglas W Oard
Gerard Salton	Vincent Quint	Mounia Lalmas
Karen Sparck Jones	Henry S Baird	Thomas S Huang
Alan F Smeaton	Dieter Merkl	James P Callan
J Stephen Downie	Ellen M Voorhees	U.S. Government
Peter Schauble	Samuel Kaski	William R Hersh
Jussi Karlgren	Ian A Macleod	Gio Wiederhold
Robert M Losee	Edward A Fox	Edward A Fox
Ian Ruthven	Richard Furuta	Craig A Knoblock
Vijay V Raghavan	Mounia Lalmas	Djoerd Hiemstra
Rong Jin	Susan T Dumais	Amanda Spink
Ophir Frieder	Leslie Lamport	Peter Ingwersen
Patrick van Bommel	Ben Shneiderman	Alan Champneys
Stephen E Robertson	Peter Ingwersen	David Hawking
John Lafferty	Airi Salminen	**Tefko Saracevic**
Rohini K Srihari	Gary Marchionini	**William S Cooper**

4. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by NSF grant # CNS-0551597. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[3] G. Mann, D. Mimno, and A. McCallum. Bibliometric impact measures leveraging topic analysis. In *JCDL 2006*, 2006.

[4] M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Lecture Notes in Physics*, 650:337–370, 2004.

[5] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR 2006*, 2006.

[1] http://www.sigir.org/awards/awards.html

[2] http://www.ukeig.org.uk/awards/tonykentstrix.html

[3] http://www.asis.org/awards/merit.htm