'Similar Researcher Search' in Academic Environments

Sujatha Das G.
Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802 USA
gsdas@cse.psu.edu

Prasenjit Mitra, C. Lee Giles
Information Sciences and Technology
Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802 USA
{pmitra,giles}@ist.psu.edu

ABSTRACT

Entity search is an emerging IR and NLP task that involves the retrieval of entities of a specific type in response to a query. We address the "similar researcher search" or the "researcher recommendation" problem, an instance of "similar entity search" for the academic domain. In response to a 'researcher name' query, the goal of a researcher recommender system is to output the list of researchers that have similar expertise as that of the queried researcher. We propose models for computing similarity between researchers based on expertise profiles extracted from their publications and academic homepages. We provide results of our models for the recommendation task on two publicly-available datasets. To the best of our knowledge, we are the first to address content-based researcher recommendation in an academic setting and demonstrate it for Computer Science via our system, `ScholarSearch`.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms

Keywords

Similar-entity Search, Recommendation

1. INTRODUCTION

Entity search and retrieval where the goal is to retrieve "objects" (such as cars, books, people) in response to user queries is an emerging research interest in the Information Retrieval community. The recent systems submitted to the entity tracks of INEX[1] and TREC[2] illustrate some approaches for facilitating entity retrieval in the general domain. In this paper, we focus on enabling entity retrieval in the academic domain where the entities of interest are researchers. We use the terms entity, expert, and researcher interchangeably in this paper. Entity retrieval has been studied in form of the "expertise search" problem in the academic domain and has been implemented in systems such as ArnetMiner[3] and Microsoft Academic Search[4]. However, to the best of our knowledge, "similar researcher search" has not been addressed via content-based approaches in the academic domain although previous work exists on predicting collaborators based on co-authorship networks [8]. Researcher recommendation differs from expertise search in that the input to the system is not a "topic query" but instead a "researcher name", and the goal is to find researchers who are similar to the queried researcher in their expertise areas. In contrast with co-author prediction, we wish to retrieve researchers who work on similar areas even if they are far apart in the co-authorship network.

As a motivating application, consider the panel selection process for a conference where the program chair wants to select a panel of reviewers for the "information extraction" track. Given an expert search system, the chair can obtain a list of "expert" recommendations for forming a panel in response to the topic query, "information extraction". On the other hand, she could use an exemplar entity in a "similar researcher search" system and search for researchers similar to "Andrew McCallum". Note that the retrieved entities in both the cases need not be the same because "Andrew McCallum" can be associated with several other expertise areas apart from "information extraction" and our recommender system seeks to retrieve researchers whose profiles are most similar to that of "Andrew McCallum" but at the same time need not be close collaborators of McCallum.

Researcher expertise profiles could be modeled using various representations. Some representations explored for the expertise search task include term vectors based on the documents authored by the researcher, probability distribution of topics s/he worked on or structural attributes describing the researcher [13, 19, 11, 3]. Irrespective of the underlying representation, for enabling "similar researcher search", we require that there exists a function that acts on two profiles and outputs a real value, $(\exists f(s_1, s_2) \rightarrow R)$, that can be used to compare the closeness between two profiles. Therefore, given the set of researchers, $E = \{e_1, e_2, \ldots e_n\}$ with profiles $(S = \{s_{e_1}, s_{e_2}, \ldots s_{e_n}\})$, an input researcher name, e_q, and a parameter k, our recommender system retrieves $E_q \subseteq E$,

[1] https://inex.mmci.uni-saarland.de/about.html
[2] http://trec.nist.gov

[3] http://arnetminer.org
[4] http://academic.research.microsoft.com

JCDL'12, June 10–14, 2012, Washington, DC, USA.

ranks them using f and outputs the top-k researchers with profiles most similar to s_{e_q}.

Contributions and Organization: We have just formally defined "researcher recommendation", an instance of "similar entity search" for the academic domain. Next, we propose models for representating researcher profiles and computing similarity with these representations (Section 2). We provide experimental evaluation for the recommendation task on two publicly-available datasets: ArnetMiner and the UvT collection[5] in Section 3. We demonstrate `ScholarSearch` (Figure 1), that implements researcher recommendation for the Computer Science domain using the data from the digital library portal, CiteSeer^{X}[6]. Finally, we summarize previous research that is closely related to our problem in Section 4 before concluding in Section 5.

2. OBTAINING SIMILAR RESEARCHERS

In Section 1, we defined the "similar researcher search" problem. Note that measuring similarity between expertise profiles presumes that we have evidence that can be used to compute similarities. For the academic domain, it is typical to measure expertise in an area in terms of a researcher's publications, descriptions of projects he or she has previously worked on, course contributions, citation information, the academic network involving a researcher, etc. This information is not easily available for all disciplines. For instance, it is more common to find research literature online for disciplines like Computer Science rather than Chemistry. Needless to say, expertise modeling depends on what evidence is available for a given discipline and various techniques are possible for extracting the same [15, 4, 20]. In the following discussion, we assume the availability of such evidence in terms of a representative document collection or at least academic homepages that concisely summarize a researcher's activities and publication information. Even if the former is not available for a discipline, previous research has illustrated techniques for obtaining researcher homepages from the web [9].

Given the set of expertise profiles for researchers, we explore the following techniques for computing similarity between two researcher profiles:

1. **Okapi BM25** (OKAPI): A researcher profile is represented using a vector corresponding to terms in a vocabulary derived based on the content associated with the researcher. Treating one profile as the query and the second as a document, ranking functions employed in IR can be used to obtain the similarity between them. Consider for instance, the Okapi BM25 ranking function widely used in various IR systems and across text collections. The similarity between two profiles is computed using the BM25 formula as follows:

$$\sum_{w \in s_1} IDF(w) * \frac{tf(w,s_2) * (k_1+1)}{tf(w,s_2) + k_1 * (1 - b + b.\frac{|D|}{avgdl}}$$

   In the above formula, $IDF(w)$ refers to the inverse document frequency of the word, a measure of rareness of the word computed as:

$$IDF(w) = log\frac{N - N(w) + 0.5}{N(w) + 0.5}.$$

   N is the total number of profiles in the collection, $N(w)$, the number of profiles containing w and $tf(w, s_2)$, the number of times, the term w appears in the profile of e_2. The parameter k_1 is typically set to a value between $[1.2, 2]$ whereas b is typically set to 0.75 in this formula in absence of other information. Additional details on this formula and parameter settings can be found in Jones, et al. [16].

2. **KL Divergence** (KLD): In this representation, a researcher profile is represented in terms of a probability distribution. For instance, given a set of documents associated with a researcher, a multinomial distribution can be fit to model the term counts in these documents and Kullback−Leibler divergence used to quantify the similarity between two probability distributions. Given θ_{s_1} and θ_{s_2}, the multinomial probability distributions associated with the profiles of researchers, e_1 and e_2 respectively, KL divergence [6] between them is given by

$$KL(\theta_{s_1}||\theta_{s_2}) = \sum_{w \in s_1} p(w|\theta_{s_1}) log \frac{p(w|\theta_{s_1})}{p(w|\theta_{s_2})}$$

3. **Probablistic Modeling** (PM): Researchers tend to work on multiple related areas and it might be more more appropriate to model their profiles as topic mixtures instead of a single multinomial distribution. Latent Dirichlet Allocation is a commonly-used topic modeling tool for unsupervised clustering of data and exploratory analysis [7]. We model the set of expertise profiles using T topics and obtain the topic distribution corresponding to each profile. The similarity between two profiles, s_1 and s_2 can now be measured in terms of the conditional probability of generating the profile s_2, from the profile $s_1, P(s_2|s_1)$. Assuming conditional independence between s_1 and s_2 given a topic and a uniform distribution on topics and entities this can be evaluated as follows

$$p(s_2|s_1) \propto \sum_{t \in T} p(s_2 \cap s_1|t) * p(t) \propto$$

$$\sum_{t \in T} p(s_2|t) p(s_1|t) \propto \sum_{t \in T} p(t|s_2) p(t|s_1)$$

   The above formulae only show terms that affect the relative ranking of profiles with respect to s_1.

4. **Trace-based Similarity** (REL): He, et al. extended van Rijsbergen's proposal to use Gleason's theorem in IR by modeling concepts as vector subspaces that are represented using density matrices (A density matrix is a symmetric, positive semi-definite matrix whose trace is 1) [18, 14]. In their formulation the density matrix for a document d, in terms of concepts c_i, $i = 1 \ldots k$ can be written as $T_d = \frac{1}{k}\sum_{i=1}^{k} c_i c_i'$ and the probability that a concept c is relevant to a document d is computed as $p_d(c) = Tr(c' T_d c)$. Based on the derivations worked out by these authors, a relevance score between two researcher profiles, s_1 and s_2 can be computed using the corresponding density matrices using the formula

$$Rel(s_1, s_2) = \frac{1}{k_1 k_2} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} (c_i.b_j)^2$$

   where $T_{s_1} = \frac{1}{k_1}\sum_{i=1}^{k_1} c_i c_i'$ and $T_{s_2} = \frac{1}{k_2}\sum_{i=1}^{k_2} b_i b_i'$. In our experiments, we computed the density matrices as

[5] http://ilk.uvt.nl/uvt-expert-collection/

[6] http://citeseerx.ist.psu.edu

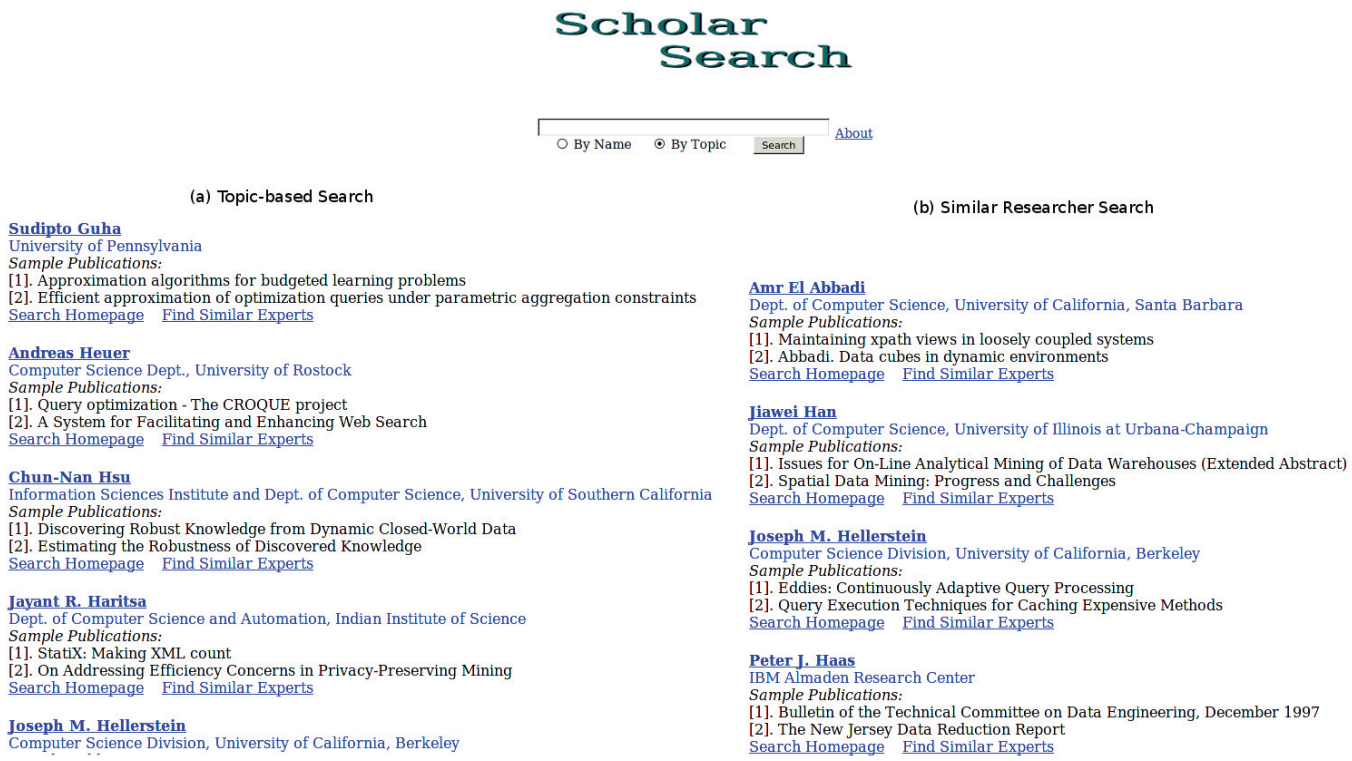

Figure 1: Demo: The ScholarSearch System

$T_s = ss'$ where s is the one-dimensional unit TFIDF vector representing a researcher's profile. More complicated techniques for setting the density matrices are possible but are a subject of future study.

3. EXPERIMENTS

To the best of our knowledge, no standard datasets exist for evaluating the academic researcher recommendation task. However, the UvT Expert collection and the datasets used in ArnetMiner [20] are publicly available for evaluating Expertise Search. These datasets contain 'topic' queries and associated with each query are manually-identified lists of researchers with expertise on the specific topic. We created datasets for evaluating the recommendation task as follows: for a given topic query, from the set of experts listed with the query, we randomly choose one of the experts as the "researcher name query". The other experts in the set comprise the similar researchers (or the 'gold' list) for this query. The lists of researchers recommended by our techniques are compared against these gold lists during evaluation. Considering only topic queries for which five or more researchers are listed with the query, we obtained a list of 163 queries for the UvT collection and 16 for ArnetMiner. The UvT collection was collected using the Webwijs system developed at Tilburg University (UvT) and contains information on UvT employees who are involved in research or teaching. The homepages, research profiles, publications and course pages of employees are included in this collection when available. Content from these sources was used to model the profiles for researchers in this dataset. For ArnetMiner, we use the document collection of CiteSeerX for modeling the expertise profiles of researchers. That is, for both the datasets, all documents associated with a given researcher are used as an aggregate document while forming the expertise profile representations (of Section 2) for that researcher.

We measure the performance of the models proposed in Section 2 on the UvT and the ArnetMiner datasets for the researcher recommendation task using average recall and Mean Reciprocal Rank (MRR) scores as a function of K (the number of retrieved results examined). Recall measures the overlap between the 'gold' lists and the retrieved lists whereas MRR indicates the rank at which the first "correct" researcher was found. As indicated in Table 1, for both the datasets, simpler models based on term vectors such as OKAPI and REL performed better than the more involved models, KLD and PM. Note that these results are indicative but an error analysis including a user study is required for a precise performance evaluation. This is because the researchers recommended by various methods may still be relevant despite not being marked as 'correct' in the 'gold' lists of our datasets. Retrieval measures that are sensitive to incomplete relevance judgements need to be studied for evaluation [1].

ScholarSearch Demo[7]: Figure 1 illustrates two modes of operation of ScholarSearch on a small subset of 15,000 authors in CiteSeer^X. Snapshot (a) denotes the results of topic-based querying for "query optimization" whereas snapshot (b) shows the results of similar-researcher search with the researcher name, "Jayant Haritsa". In this system, topic-based expertise search was implemented using the models proposed in Demartini, et al. [12]. ScholarSearch also provides the "homepage search" functionality (not shown in the

[7]will be publicly-available soon from the first author's homepage

	ArnetMiner			UvT		
	K=10	K=30	K=100	K=10	K=30	K=100
OKAPI	**0.0266**, **0.2569**	**0.0461**, **0.2662**	**0.0965**, **0.2682**	0.1175, 0.2857	0.1828, 0.2945	0.3069, 0.2979
KLD	0.0034, 0.0625	0.0034, 0.0625	0.0119, 0.0665	0.0701, 0.1555	0.1128, 0.1605	0.2169, 0.1654
PM	0.0062, 0.1063	0.0130, 0.1104	0.0224, 0.1125	0.0907, 0.1591	0.2177, 0.1746	0.4239, 0.1782
REL	0.0203, 0.2188	0.0360, 0.2267	0.0892, 0.2303	**0.1787**, **0.3006**	**0.2997**, **0.3112**	**0.4774**, **0.3131**

Table 1: Average Recall and MRR values on the ArnetMiner and UvT datasets

figure) using the ranking function previously proposed by us [10]. The "Find Similar Experts" option implements the researcher recommendation task described in this paper, using the **REL** (Section 2) model. We provide the top-5 recommendations retrieved by our system in response to a few researcher names in Table 2. As the anecdotal evidence indicates ScholarSearch predictions for "similar researchers" in response to the queried researcher's name are quite reasonable.

Q1. Andrew McCallum	**Q2. Avi Wigderson**
James Allan	Frederic Green
Lise Getoor	Vikraman Arvind
Fernando C. N. Pereira	Eric Allender
Dan Roth	Sanjeev Arora
Thomas G. Dietterich	Umesh V. Vazirani
Q3. Christopher Manning	**Q4. W. Bruce Croft**
Eugene Charniak	Omar Alonso
Amy Weinberg	Dawn J. Lawrie
Kathleen McKeown	Dell Zhang
Nicholas Bambos	Michael S. Lew
Mark Liberman	Jesus Vilares Ferro

Table 2: Top-5 recommendations by ScholarSearch for sample queries, Q1-Q4.

4. RELATED WORK

The list-completion tasks in TREC and INEX address the similar-entity finding task in the general domain. The proceedings of these competitions discuss various systems that were designed to handle this task. In contrast to our problem, the input queries in these systems, include a query topic description with examples of entities. The participating systems need to extract the relation between the example entities and the topic description and propose entities that hold a similar relation with the topic description, as part of the answer. Similar expert finding was addressed by Balog, et al. on the TREC data using the relations a candidate expert has with other experts, documents and terms [5]. Hofmann, et al. considered the contextual factors such as organizational setup and combined them with content-based retrieval scores to find similar experts within an organization [15]. Although we could not find previous work on content-based similar-entity finding in academic disciplines, previous work exists for predicting researchers to collaborate with. Chen, et al. presented CollabSeer that uses the structure of the co-author network to predict research collaborators [8]. Xu, et al [21] use a two-layer network model that combines co-author network and researcher-concept network for making researcher recommendations. However, our approach targets the prediction of researchers with similar expertise profiles based on content they generate and not on their distance in the co-authorship graph. Several models also address the closely-related task of expert search/ranking given a topic query both in academic domains and enterprises. Typically, a document collection available in a domain is used to infer the expertise of an author based on the authorship information [2, 17, 13]. In some cases, the underlying connections between documents, researchers and other entities can be explored via graph-based models [19, 11].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we formulated the researcher recommendation problem in academic environments. We discussed several techniques for representing expertise profiles based on the available evidence and proposed models for computing similarity between two profiles. We evaluated our proposed techniques on two publicly-available datasets and showed the viability and usability of researcher recommendation via our demo system, `ScholarSearch`. We are currently focusing on improving the accuracy as well as the response time of our recommendation system. To this end, we are exploring techniques for a more accurate representation for researcher profiles, not just in terms of the documents but including also the underlying academic network between researchers and metadata information such as university affiliations.

Acknowledgments: We acknowledge support from NSF (Grant Nos. 0845487, 0949891 and 0958143) and DTRA (Grant No. HDTRA1-09-1-0054) for this work.

6. REFERENCES

[1] P. Ahlgren and L. Grönqvist. Evaluation of retrieval effectiveness with incomplete relevance data: Theoretical and experimental comparison of three measures. *Inf. Process. Manage.*, 2008.
[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, 2006.
[3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR*, 2007.
[4] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *IJCAI*, 2007.
[5] K. Balog and M. de Rijke. Finding similar experts. In *SIGIR*, 2007.
[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2007.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.
[8] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: a search engine for collaboration discovery. In *JCDL*, 2011.
[9] S. Das, C. L. Giles, P. Mitra, and C. Caragea. On identifying academic homepages for digital libraries. In *JCDL*, 2011.
[10] S. Das, P. Mitra, and C. L. Giles. Learning to rank homepages for researcher name queries. In *EOS Workshop at SIGIR*, 2011.
[11] S. Das, P. Mitra, and C. L. Giles. Ranking authors in digital libraries. In *JCDL*, 2011.
[12] G. Demartini, J. Gaugaz, and W. Nejdl. A vector space model for ranking entities and its application to expert search. In *ECIR*, 2009.
[13] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *ICDM*, 2008.
[14] Q. He, J. Pei, D. Kifer, P. Mitra, and C. L. Giles. Context-aware citation recommendation. In *WWW*, 2010.
[15] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke. Contextual factors for finding similar experts. *JASIST*, 2010.
[16] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 2000.
[17] D. Mimno and A. McCallum. Mining a digital library for influential authors. In *JCDL*, 2007.
[18] C. J. v. Rijsbergen. *The Geometry of Information Retrieval*. 2004.
[19] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM*, 2008.
[20] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.
[21] Y. Xu, J. Hao, R. Y. Lau, J. Ma, W. Xu, and D. Zhao. A personalized researcher recommendation approach in academic contexts: Combining social networks and semantic concepts analysis. In *PACIS*, 2010.