On Information-Theoretic Document-Person Associations for Expert Search in Academia

Vitor Mangaravite
mangaravite@dcc.ufmg.br

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil

ABSTRACT

State-of-the-art expert search approaches rely on document-person associations to infer the expertise of a candidate person for a given query. Such associations have traditionally been modeled as boolean variables, indicating whether or not a candidate authored a document, and further normalized to penalize prolific authorships. In this paper, we address expert search in academia, where the authorship of a document can be determined with reasonable certainty. In contrast to traditional approaches, we propose to model associations as non-boolean variables, reflecting the probability that a document is *informative* of the expertise of a candidate. Moreover, we introduce an alternative normalization scheme that measures how *discriminative* a particular document-person association is in light of all associations involving either the document or the person. Through a large-scale user study with academic experts from several areas of knowledge, we demonstrate the suitability of the proposed association and normalization schemes to improve the effectiveness of a state-of-the-art expert search approach.

CCS Concepts

•**Information systems** → **Expert search;**

Keywords

Academic search; expertise retrieval

1. INTRODUCTION

Users search for an expert whenever they need proficient knowledge on a given topic [7]. For instance, an expert on "*information retrieval*" could be searched for providing consultancy or for being recruited by a search company. Expert search has received considerable attention from the information retrieval community over the past decade, with a particular focus on finding experts within an enterprise organization [2, 8, 10, 19]. Several expert search approaches have been proposed that attempt to model the expertise of candidate persons and their relevance given a user's query (e.g., [3, 12, 15]). In common, all of these approaches rely on some form of association between people and documents in order to model the expertise profile of each candidate [7].

Document-person associations are commonly modeled as boolean variables, indicating whether or not a person has authored or is mentioned in a document. Given the ambiguous nature of such associations in typical enterprise collections [1, 10], attempts to model non-boolean associations have been made that reflect the confidence that the right person has been identified (e.g., based upon the frequency of occurrence of the person's name in the document [6] or its occurrence in proximity to the query terms [20]). In addition, to limit the impact of false positives, the inferred associations are typically normalized to penalize prolific authorships. In contrast, many other expert search scenarios provide unambiguous associations, such as email-sender in email corpora [5] and paper-author in academic corpora [4, 11], which preclude the need for authorship inference.

In this paper, we address expert search in academia. In this scenario, the authorship of a document can be determined with reasonable certainty, by leveraging metadata associated with individual publication records. As a result, rather than attempting to infer the authorship of a document, we propose to model associations as non-boolean variables reflecting the probability that the document is *informative* of the expertise of a candidate. Moreover, because authorships in this scenario are unambiguous, penalizing documents with many associated authors or authors with many associated documents becomes arguably counterintuitive. Therefore, we further introduce an alternative normalization scheme that measures how *discriminative* a particular document-person association is in light of all associations involving either the document or the person. Through a large-scale user study with academic experts from several areas of knowledge, we demonstrate the suitability of the proposed association and normalization models to improve the effectiveness of a state-of-the-art expert search approach. To the best of our knowledge, this is the first attempt to infer the strength of document-person associations beyond authorship attribution for expert search in academia.

In the remainder of this paper, Section 2 discusses related work on expert search and association models. Section 3 introduces our information-theoretic models for weighting document-person associations. Section 4 describes the setup and the results of the empirical evaluation of our proposed models. Finally, Section 5 presents our conclusions.

SIGIR '16, July 17–21, 2016, Pisa, Italy.

DOI: http://dx.doi.org/10.1145/2911451.2914751

2. RELATED WORK

Expert search has been the subject of intense research over the past decade, particularly with the introduction of an expert search task at the TREC 2005 Enterprise track [10]. For instance, Balog et al. [3] introduced two alternative generative probabilistic approaches for expert search, which estimate the likelihood that a given query is generated by each candidate expert. Their Model 1 performs this estimation directly, by relying on the candidate's language model built from his or her associated documents. In turn, their Model 2 assumes a generative process in which candidates generate documents, which themselves generate the query. Discriminative probabilistic models have also been proposed, which attempt to estimate the relevance of a given query-candidate pair directly from training data [12]. Other prominent approaches include graph-based models [18], which perform inference on an expertise graph built from document-person associations, and voting models [15], which aggregate the query-biased document relevance estimates into relevance estimates for the associated candidate experts.

A common characteristic of most existing expert search approaches is their reliance on document-person associations. In particular, almost all approaches consider boolean associations, indicating whether or not a candidate has authored a particular document. However, in a typical enterprise setting, document-person associations can be ambiguous. For instance, candidate names may be mentioned in different parts of a meeting minute without any explicit authorship indication. To overcome such an ambiguity, non-boolean association models have been proposed to estimate the confidence that the correct candidate has been identified. For instance, Balog and de Rijke [6] proposed to weight document-person associations based on the Kullback-Leibler divergence between the document and the candidate's language models built using only the candidate's name. Other association models also estimated the distance between the candidate's name and the query terms to improve the identification of authorships related to the query topic [20].

In contrast to the aforementioned approaches, we tackle expert search in academia, a domain where authorships can be determined with reasonable certainty via publication metadata records uniquely associated with each candidate. As a result, we propose to model associations as non-boolean variables aimed to convey the *informativeness* of each document for the expertise profile of a candidate. In a similar spirit, Macdonald and Ounis [14] proposed to estimate the extent to which a document is related to the core interests of each candidate associated with it. To this end, they clustered each candidate's profile and ranked the resulting clusters in decreasing order of their size, weighting a document-person association by the inverse of the rank of the cluster that contained the document. Our proposed approach is simpler, and relies on a standard information-theoretic measure of the informativeness of the document for the candidate's expertise. The cluster-based approach of Macdonald and Ounis [14] is included as a baseline in our experiments.

Another aspect of document-person associations that is relevant to our proposal is normalization. For instance, both generative models proposed by Balog et al. [3] rely on a normalization component to estimate the probability $p(d|e)$ that document d is associated with candidate e. This estimation can be either document- or candidate-centric, depending on whether the association weight is normalized by the sum of the weights of all associations related to the document or the candidate, respectively. Normalizing associations may be useful even for non-probabilistic approaches, to counter the noise introduced by spurious associations resulting from false-positive authorship attributions. With this in mind, Macdonald and Ounis [16] proposed to normalize associations by the length of the candidate's profile, measured in either number of terms or number of documents. Given the absence of spurious associations in our scenario, we propose a normalization scheme that does not penalize prolific candidates, but instead measures how discriminative each association is for each document or candidate. The normalization schemes of Balog et al. [3] and Macdonald and Ounis [16] are used as baselines in our investigation.

3. INFORMATION-THEORETIC DOCUMENT-PERSON ASSOCIATIONS

A key element of any expert search approach is its ability to represent the expertise of each candidate expert. As discussed in the previous sections, most expert search approaches in the literature rely on explicit document-person associations to model a candidate's expertise profile. Without loss of generality, we can formalize the association $f(d, e)$ between document d and candidate e according to:

$$f(d, e) = \psi(\rho(d, e)), \tag{1}$$

where $\rho(d, e)$ and $\psi(\bullet)$ denote *association* and *normalization* schemes for the association between document d and candidate e, respectively. The majority of the expert search approaches in the literature rely on a boolean association scheme, which assigns a constant value $\rho(d, e) = 1$ for all existing associations. In turn, the most straightforward normalization scheme simply divides the association $\rho(d, e)$ by the sum of all document or candidate associations, such that $\psi(\bullet) \equiv \bullet / \sum_{e'} \rho(d, e')$ or $\psi(\bullet) \equiv \bullet / \sum_{d'} \rho(d', e)$ [3].

In the following, we introduce novel information-theoretic association and normalization schemes for expert search in academia. The proposed schemes exploit the unambiguous nature of document-person associations leveraged from publication records in order to better quantify the informativeness of each association for a candidate's expertise.

3.1 Association Scheme

Our proposed association scheme aims to weight a given document-person association based on how informative the document is of the expertise of the candidate. To this end, we rely on an information-theoretic measure of the distance between the language use in the document and in the entire profile of the candidate. Precisely, our proposed association scheme $\rho_H(d, e)$ can be instantiated as follows:

$$\begin{aligned} \rho_H(d, e) &= H(\theta_e, \theta_d) \\ &= -\sum_t p(t|\theta_e) \log p(t|\theta_d), \end{aligned} \tag{2}$$

where $H(\theta_e, \theta_d)$ is the cross-entropy between the candidate language model θ_e—built by concatenating all documents associated with candidate e—and the document language model θ_d. It can be shown that this formulation is equivalent to estimating the likelihood of generating the candidate language model θ_e given the document language model θ_d [13]. It is also equivalent to the standard Kullback-Leibler divergence between the two models minus the entropy $H(\theta_e)$

of the candidate model θ_e, which is discarded here as it is the same for all documents d. Based upon the latter observation, our proposed formulation can also be seen as a generalization of the association scheme proposed by Balog et al. [6] and discussed in Section 2. While their approach is based on language models built from candidates' names, we evaluate our approach using multiple textual representations for the document and candidate models. Finally, while alternative formulations (e.g., based on non-textual features) are possible, we leave their investigation for future research.

3.2 Normalization Schemes

Existing expert search approaches linearly normalize association weights either by the sum of all related association weights [3] or by the length of the candidate's profile [16]. As discussed in Section 2, these approaches have been shown to be effective, particularly in the enterprise domain, where ambiguous associations may spuriously compromise the estimation of a candidate's expertise. In the academic domain, where associations are generally unambiguous, such linear normalization approaches may harshly penalize prolific candidates. To counter this limitation, we propose two alternative normalization schemes aimed to measure the discriminativeness of each association. In particular, our soft document-centric (SDC) normalization scheme is given by:

$$\psi_{SDC}(\bullet) \equiv \log \frac{\left(\sum_{e'} \rho(d, e')\right)^{\alpha}}{\bullet}, \quad (3)$$

where the summation in the numerator comprises the weights of all other associations related to target document d. Our soft candidate-centric (SCC) normalization scheme can be defined analogously with respect to target candidate e:

$$\psi_{SCC}(\bullet) \equiv \log \frac{\left(\sum_{d'} \rho(d', e)\right)^{\alpha}}{\bullet}. \quad (4)$$

In both the SDC and SCC schemes, the logarithm provides for a softer normalization compared to existing approaches from the literature, with parameter α in the numerators controlling the intensity of the normalization—the larger α, the softer the normalization. With $\alpha = 1$, both normalization schemes reduce to the information-theoretic concept of self-information [9], which quantifies the improbability of occurrence of a particular association given all other associations related to the same document or candidate.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate our information-theoretic association models for expert search in academia. In particular, we aim to answer two research questions:

Q1. How effective is our proposed association scheme?

Q2. How effective are our proposed normalization schemes?

In the following, we describe the setup and discuss the results of our empirical investigations.

4.1 Experimental Setup

Our evaluation is based on the Lattes Expertise Retrieval (LExR) test collection [17], a publicly available test collection built on top of the Lattes platform,[1] an internationally renowned initiative for managing information about science, technology, and innovation for individual researchers and research institutions in Brazil. The LExR test collection comprises metadata records for 11,942,014 scientific publications associated with 206,697 candidate experts from all areas of knowledge working in multiple Brazilian research institutions spread all over the country. Moreover, it includes 235 queries suggested by real experts who judged one another on a graded scale. Grades 0, 1, 2, and 3 indicate an unknowledgeable, somewhat knowledgeable, very knowledgeable, and expert person on the query topic, respectively.

As baseline association schemes, we consider a standard boolean scheme (B), as well as non-boolean schemes based on the divergence (KL) of document and candidate language models built from each candidate's name [6] and on a clustering (CL) of each candidate's profile [14]. As baseline normalization schemes, we consider the standard document-centric (DC) and candidate-centric (CC) schemes [3] as well as schemes aimed at profile length normalization based on terms (Nt) and documents (Nd). All association and normalization schemes are deployed on top of Model 2 [3] as a representative of state-of-the-art expert search approaches. Model 2 is set to operate with the top 1,000 documents retrieved by a standard language model with Dirichlet smoothing with parameter $\mu = 2,000$. We index the title, keywords, abstract, and author names as separate document fields after removing stop words and applying no stemming. All indexing and retrieval operations use Apache Lucene 5.3.[2]

Our proposed association scheme in Equation (2) is deployed with document and candidate language models built using the abstract and author name fields of each document. Other textual representations showed similar effectiveness. In turn, our normalization schemes in Equations (3) and (4) are deployed with $\alpha = 2$, which showed marginal improvements compared to the standard setting of $\alpha = 1$. A full parameter sensitivity analysis is left for future research.

4.2 Experimental Results

Table 1 shows the retrieval effectiveness of several association (ρ) and normalization (ψ) schemes in terms of normalized discounted cumulative gain (nDCG@10), precision (P@10), and mean reciprocal rank (MRR). Our proposed association (ρ_H, Equation (2)) and normalization (ψ_{SDC} and ψ_{SCC}, Equations (3)-(4)) schemes are highlighted in gray. Statistical significance is verified using a paired t-test with $p < 0.01$. Superscript and subscript triangles (▲) denote significant improvements against the boolean association scheme (ρ_B) and the standard document- (ψ_{DC}) or candidate-centric (ψ_{CC}) normalization schemes, respectively.

From Table 1, in order to address research question Q1, we first analyze the effectiveness of our proposed association scheme. In particular, when using a standard document-centric normalization scheme (ψ_{DC}), our association scheme ρ_H consistently outperforms all other association schemes (ρ_B, ρ_{CL}, and ρ_{KL}) with respect to all evaluation metrics, with significant improvements compared to the strongest boolean association baseline (ρ_B). On the other hand, when using the standard candidate-centric normalization (ψ_{CC}), both the boolean association baseline as well as our proposed association scheme underperform, probably because of the harsh penalization applied to prolific candidates. Recalling research question Q1, these results attest the effectiveness

[1] http://lattes.cnpq.br/

[2] http://lucene.apache.org/

	nDCG	P@10	MRR	nDCG	P@10	MRR
ρ_{CL}	0.022	0.014	0.060	–	–	–
ρ_{KL}	0.135	0.082	0.240	–	–	–
	ψ_{Nt}			ψ_{Nd}		
ρ_B	0.012	0.009	0.039	0.061	0.008	0.028
	ψ_{DC}			ψ_{CC}		
ρ_B	0.133	0.079	0.254	0.004	0.004	0.018
ρ_H	$0.146^{\blacktriangle}$	$0.088^{\blacktriangle}$	$0.279^{\blacktriangle}$	0.005	0.004	0.020
	ψ_{SDC}			ψ_{SCC}		
ρ_B	$0.164_{\blacktriangle}$	$0.096_{\blacktriangle}$	$0.294_{\blacktriangle}$	$0.167_{\blacktriangle}$	$0.102_{\blacktriangle}$	$0.295_{\blacktriangle}$
ρ_H	$0.164_{\blacktriangle}$	$0.099_{\blacktriangle}$	$0.291_{\blacktriangle}$	$0.168_{\blacktriangle}$	$0.103_{\blacktriangle}$	$0.293_{\blacktriangle}$

Table 1: Retrieval effectiveness of several association (ρ) and normalization (ψ) schemes. Our proposed weighting schemes are highlighted in gray.

of our proposed information-theoretic association scheme, particularly under a document-centric normalization.

To address research question Q2, we contrast our soft normalization schemes (ψ_{SDC} and ψ_{SCC}) to their standard counterparts (ψ_{DC} and ψ_{CC}) and the two profile length normalization baselines (ψ_{Nt} and ψ_{Nd}). From Table 1, we observe that both of our proposed normalization schemes substantially outperform all baselines, with significant improvements compared to the standard document- and candidate-centric baselines with respect to all metrics. Recalling question Q2, these observations confirm the effectiveness of our proposed soft normalization schemes. Indeed, substantial improvements can be observed even when normalizations are candidate-centric, which further emphasizes the benefits of avoiding a harsh penalization of prolific candidates. On the other hand, the benefits of our improved association scheme ρ_H over boolean associations seem to be offset, which suggests further investigations towards an improved combination of association and normalization schemes.

5. CONCLUSIONS

We have proposed novel information-theoretic models to weight document-person associations for expert search in academia. Unlike existing approaches, which attempt to ascertain the authorship of a document, we proposed an association scheme aimed to estimate the extent to which the document is informative of the expertise of each candidate associated with it. Likewise, rather than penalizing prolific documents or candidates to avoid noisy estimations of expertise, we proposed soft normalization schemes that seek to infer how discriminative each association is. We evaluated our proposed association and normalization schemes using a publicly available test collection for expertise retrieval in academia encompassing candidate experts from multiple organizations and with expertise in a range of diverse knowledge areas. Our results demonstrated the effectiveness of the proposed weighting schemes, with substantial and statistically significant improvements over several baselines from the literature. As a direction for future research, we plan to extend our current investigation to non-textual association schemes, exploiting temporal and social signals of expertise.

Acknowledgments

This work was partially funded by projects InWeb (MCT/CNPq 573871/2008-6) and MASWeb (FAPEMIG/PRONEX APQ-01400-14), and by the authors' individual grants from CNPq and FAPEMIG.

6. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

[2] P. Bailey, A. P. de Vries, N. Craswell, and I. Soboroff. Overview of the TREC 2007 Enterprise track. In *Proc. of TREC*, 2007.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. of SIGIR*, pages 43–50, 2006.

[4] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *Proc. of SIGIR*, pages 551–558, 2007.

[5] K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *Proc. of WWW*, pages 1035–1036, 2006.

[6] K. Balog and M. De Rijke. Associating people and documents. In *Proc. of ECIR*, pages 296–308, 2008.

[7] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Found. Trends Inf. Retr.*, 6(2–3):127–256, 2012.

[8] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. P. de Vries, and P. Bailey. Overview of the TREC 2008 Enterprise track. In *Proc. of TREC*, 2008.

[9] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[10] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise track. In *Proc. of TREC*, 2005.

[11] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on DBLP bibliography data. In *Proc. of ICDM*, pages 163–172, 2008.

[12] Y. Fang, L. Si, and A. P. Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proc. of SIGIR*, pages 683–690, 2010.

[13] V. Lavrenko and W. B. Croft. *Relevance Models in Information Retrieval*, chapter 2. 2003.

[14] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *Proc. of ECIR*, pages 283–295, 2008.

[15] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of CIKM*, pages 387–396, 2006.

[16] C. Macdonald and I. Ounis. Searching for expertise: experiments with the Voting Model. *Comput. J.*, 52(7):729–748, 2009.

[17] V. Mangaravite, R. L. T. Santos, I. S. Ribeiro, M. A. Gonçalves, and A. H. F. Laender. The LExR collection for expertise retrieval in academia. In *Proc. of SIGIR*, 2016.

[18] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *Proc. of CIKM*, pages 1133–1142, 2008.

[19] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise track. In *Proc. of TREC*, 2006.

[20] J. Zhu, D. Song, and S. Rüger. Integrating multiple windows and document features for expert finding. *J. Am. Soc. Inf. Sci. Technol.*, 60(4):694–715, 2009.