A unified approach for learning expertise and authority in digital libraries

Baptiste De La Robertie, Liana Ermakova, Yoann Pitarch, Atsuhiro Takasu, Olivier Teste

A Unified Approach for Learning Expertise and Authority in Digital Libraries

B. de La Robertie[1(✉)], L. Ermakova[2,3], Y. Pitarch[1], A. Takasu[4], and O. Teste[1]

[1] Université de Toulouse, IRIT UMR5505, 31071 Toulouse, France
baptiste.robertie@gmail.com, {o.teste,y.pitarch}@irit.fr
[2] Université de Lorraine, Nancy, France
liana.ermakova.@irit.fr
[3] LISIS, Université de Paris-Est Marne-la-Vallée, Champs-sur-Marne, France
[4] National Institute of Informatics, 2-1-2 Hitotsunashi, Chiyoda, Tokyo, Japan
takasu@nii.ac.jp

Abstract. Managing individual expertise is a major concern within any industrial-wide organization. If previous works have extensively studied the related expertise and authority profiling issues, they assume a semantic independence of these two key concepts. In digital libraries, state-of-the-art models generally summarize the researchers' profile by using solely textual information. Consequently, authors with a large amount of publications are mechanically fostered to the detriment of less prolific ones with probably higher expertise. To overcome this drawback we propose to merge the two representations of expertise and authority and balance the results by capturing a mutual reinforcement principle between these two notions. Based on a graph representation of the library, the expert profiling task is formulated as an optimization problem where latent expertise and authority representations are learned simultaneously, unbiasing the expertise scores of individuals with a large amount of publications. The proposal is instanciated on a public scientific bibliographic dataset where researchers' publications are considered as a source of evidence of individuals' expertise and citation relations as a source of authoritative signals. Results from our experiments conducted over the Microsoft Academic Search database demonstrate significant efficiency improvement in comparison with state-of-the-art models for the expert retrieval task.

Keywords: Expert finding · Link analysis · Optimization · Digital libraries

1 Introduction

Keeping track and managing individuals' expertise in industrial-wide organizations or public scientific repositories is a major concern. Motivated by expertise capitalization, skill mining, or knowledge sharing purposes, strong interests on

the expert finding task rapidly spawned both private and public researches [25]. For example, the Experscape platform[1], by mining the US National Library of Medicine and the National Institutes of Health databases[2], provides search functionalities to seek experts according to 26,000 topics (e.g., Alzheimer Disease, Arthritis, Brain Tumor) and geographic features (country, region, city, and institution). The system AMiner[3], resting on DBLP[4] and ACM[5], also provides search functionalities for the Computer Science field and capitalizes more than 100 million researchers and 200 million publications. Microsoft Academic Search[6] and more recently ResearchGate[7] also constitute popular examples exploiting digital libraries for profiling and discovering goals.

While expert profiling and retrieval attract significant interest by the scientific community, unified approaches that consider both expertise and quality models receive too little attention. Indeed, state-of-the-art models generally formulate the expert finding problem as a summarization task where text data, essentially associated to individuals, are used to model knowledge and expertise [1,6,20,22]. Intranet documents, reports, project descriptions, mails, or publications are used as a source of information whereas tags, key words, or flat topics are extracted to link knowledge and experts [6]. *In fine*, candidates are then ranked according to the probability of being an expert given a particular topic. The underlying matching process, generally based on standard information retrieval techniques, ignores quality or authoritative criteria. Therefore, authors with larger amounts of productions are promoted, biasing the final ranking over the candidates.

To illustrate this downside, let us consider the following example. Let $\mathcal{R} = \{r_1, r_2\}$ and $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$ be 2 sets of 2 researchers and 5 articles respectively. The authoring relation between researchers and articles is given in Fig. 1. Let θ_1 and θ_2 be the profiles associated to the researchers r_1 and r_2 respectively. We consider a language model formalism for summarization. Given a query q, researchers are ranked according to the probability $p(q|\theta_i) = \prod_{w \in q} p(w|\theta_i)$. Using Bayes' rules, it holds $p(w|\theta_i) = \sum_{a_j \in \mathcal{A}} p(w|a_j)p(r_i|a_j)$, making the value of $p(w|\theta_i)$ increasing with the number articles authored by a researcher r_i. For example, given a topic query $q = \{w_1\}$ where the term probability $p(w_1|a_j)$ is the same for all articles a_j (see Fig. 1), $p(q|\theta_1) = 0.4 < p(q|\theta_2) = 0.6$. Thus, the researcher r_2 is promoted with regard to topic w_1. However, if the articles authored by r_1 are much more cited than those authored by r_2, one will probably rank researcher r_1 higher than r_2. This example motivates the need of considering quality or authority signal in a profile summarization task.

[1] http://expertscape.com.
[2] https://www.ncbi.nlm.nih.gov.
[3] https://aminer.org/.
[4] http://dblp.uni-trier.de/.
[5] http://dl.acm.org/.
[6] http://academic.research.microsoft.com/.
[7] https://www.researchgate.net.

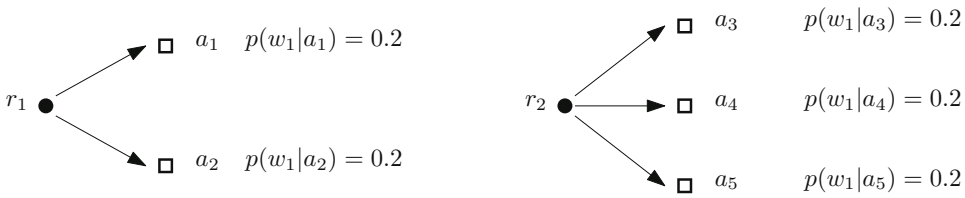

Fig. 1. Illustration of the drawback of state-of-the-art expert profiling methods. Without considering any quality or authority signal, researcher r_2 will be ranked higher than researcher r_1 for similar topic queries.

In this work, we tackle this drawback by assuming that expertise and authority influence each other. We assume that (1) experts are sources of knowledge (associated publications contain proofs of expertise), (2) experts are authoritative (associated relations contain proofs of authority) and (3) these two components, being two sides of the same coin, should have a common representation. To capture this mutual reinforcement principle, we formulate the expert profiling task as an optimization problem where both authority and expertise vectors are unified and simultaneously learned. As confirmed by the experiments, such a representation for expert profiling significantly improves the expert finding phase. To summarize, our contributions are as follows:

1. We provide a unified model capturing both individuals' expertise and authority based on an heterogeneous graph representation of digital libraries;
2. We formulate the expert profiling task as an optimization problem learning both latent topics and authoritative signals in a single process;
3. We conduct experiments over a representative subset of the Microsoft Academic Search (MAS) database and show a significant improvement as compared to state-of-the-art methods.

The rest of the paper is organized as follows: Sect. 2 discusses the related work. Section 3 formally describes our model. Section 4 discusses the experiments. Finally, concluding remarks are drawn in Sect. 5.

2 Related Work

Our model relates to both expertise and authority fields. We first provide an overview of these two research topics and then motivate the need of a unified approach for modelling authority and expertise using a single representation.

Expertise profiling and retrieval models. Historical approaches related to expert finding manually store individuals' skills in knowledge bases [7]. The distinction between the representation of knowledge and data is manually made on the basis of reports, scientific articles or employee pages but presents considerable maintenance costs. Craswell et al. [6] first propose an automatic solution, assimilating an employe's profile to the concatenate list of his/her related documents. Thus, given a topic query, standard information retrieval techniques are

used to retrieve the top-n experts. State-of-the-art models generally make use of language or topic models to summarize an individual profile [3]. In this category, extensive works have been done by Balog et al. [1,2,19] by proposing a generative probabilistic modeling framework for expert profiling. Standard Information Retrieval techniques are adapted for that task, estimating a probability of a candidate being an expert in a particular topic. For a given topic query q, candidates are ranked according to the probability $p(q|\theta_{ca})$ where the representation of a candidate θ_{ca} is generally performed using a multinomial probability distribution over a vocabulary (i.e., $p(q|\theta_{ca}) = \prod_{t \in q} p(t|\theta_{ca})^{n(t,q)}$). In [20], the expert profiling task is formulated as a tagging problem where features extracted from various sources are used to model an employee. In particular, authored enterprise documents, discussion lists, and enterprise search click-through data are used to learn a tag probability of being a good descriptor for a particular employee. In [21], the web user profiling problem is tackled on the basis of topic modelling, without considering authority signal. In all these previous works, only the textual content is used for expert profiling which constitutes the introduced major drawback. Yang et al. [24] integrate authoritative features using the PageRank scores of researchers. Nevertheless pre-computed scores and some other language model features are aggregated *a postiori*, feeding a feature vector for training. The proposition cannot capture any cyclic relation between the two concepts. Deng et al. [9] construct a weighted language model to take into consideration not only the relevance between a query and documents but also the importance of the documents. Only the number of citations of an article is integrated as a prior probability. Thus, the notions of authority and expertise in the literature are generally separated and do not influence each other.

Graph-based authoritative models. In organizational networks, graph-based models, largely based on random walk [23], are widely used to estimate individual authority. In this field, extensive researches have demonstrated strong correlations between centrality and authority [10,16,26,27]. The famous PageRank algorithm proposed by L. Page et al. [18] and later the Topic-sensitive Pagerank [11] have proven the value of the citation graph for web pages. Campbell et al. [5] exploit network patterns in email communication graphs to discover experts and show that a HITS-like algorithm [14] performs better than content-based approaches for the expert finding task. The co-author graph on Wikipedia has demonstrated to carry out authority signals and help in identifying authoritative users producing high quality content [8]. Jurczyk et al. [13] also make used of a HITS-based algorithm to estimate the authority of Question and Answering platforms' members and confirm the robustness of such approach. Finally, Takasu et al. [12] employ both co-author and citation graphs to discriminate researchers' importance rating. State-of-the-art approaches demonstrate the efficiency of graph-based authority models but also the lack of unified approaches considering expertise.

Discussion. Propositions considering both expertise and authority signals have received too little attention. Unified approaches widely compute two

representations then aggregate them *a posteriori* preventing from capturing a mutual reinforcement principle. Unlike previous well-established methods, we propose to formulate the expert profiling problem as a summarization task where both expertise and authority concepts are merged into a single representation. An individual is considered as an expert not only if he/she authors some articles in a particular field but also if the authored articles are credited by the community. Moreover, the unified representation enables us to strengthen the scores of poorly represented dimensions of a researcher's profile who would have authored few but highly cited articles. Conversely, this unified representation enables to balance the scores of over-represented dimensions of a researcher's profile who would have abundantly written poor quality articles. To the best of our knowledge, our proposition is the first approach connecting the two key concepts for the expertise retrieval task.

3 Model

A digital library can naturally be represented by an heterogeneous directed graph, denoted by G, where the sets of nodes U correspond to the different entities in the library and the sets of edges V to the different relations defined by the platform. In this work, G encodes the sets of articles, researchers and words in addition to the authoring and citing relations. Unlike state-of-the-art models, we assume that individuals' expertise and authority share a common representation in $\mathbb{R}^K$, encoding to what extent a researcher is an expert *and* he/she is authoritative in a particular field. We suppose that the content of the articles contains proof of expertise and the relative locations of the articles in the citation graph constitute proof of quality of the articles. Thus, we propose to compute the profile of a researcher as an aggregation of the estimated expertise and quality of the authored articles. Section 3.1 introduces the general notations for representing a digital library. Section 3.2 details the proposed unified representation for capturing the cyclical relation between expertise and authority. Finally, the objective function to learn expertise and authority simultaneously is detailed in Sect. 3.3.

3.1 Platform Representation

Let $\mathcal{R} = \{r_i\}_{1 \leq i \leq N}$, $\mathcal{A} = \{a_j\}_{1 \leq j \leq M}$ and $\mathcal{W} = \{w_s\}_{1 \leq s \leq W}$ be the sets of researchers, articles and words respectively. We define the heterogeneous graph $G = (U, V)$ over the set of nodes $U = \mathcal{R} \cup \mathcal{A} \cup \mathcal{W}$ and relations $V = V_{RA} \cup V_{AA} \cup V_{AW}$. In particular, V_{RA} is an authoring relation associating each researcher to the articles he/she authored. V_{AA} is a citing relation. Finally, V_{AW} associates each article to the set of words it contains. Corresponding adjency matrices are denoted by X_{RA}, X_{AA} and X_{AW} respectively. Note that X_{RA} and X_{AA} are binary matrices (i.e., $X_{RA}(i, j) = 1$ if researcher r_i has authored article a_j, 0 otherwise). In this work, X_{AW} contains TF-IDF weights. Notations are summarized in Fig. 2. Latent representations of expertise and authority are detailed in the next section.

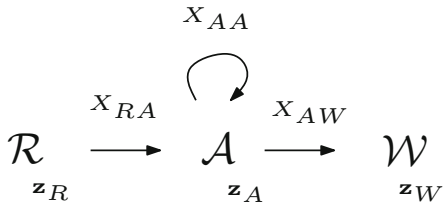

Fig. 2. Graphical representation of a digital library.

3.2 Encoding Expertise and Authority

We propose to represent both expertise and authority in a single vector in $\mathbb{R}^K$ where the k-th dimension is aimed to estimate both expertise and authority of a particular entity in G for a latent topic k. Figure 3 illustrates the proposed formulation for a particular researcher and 4 topics. The mutual reinforcement principle between the expertise of the researcher and his/her authority in topic 3 figures out by dotted arrows. In order to unbias expertise scores associated to over-represented or under-represented dimensions, we propose to merge these two vectors. The proposed unified representation estimates without distinction both expertise and authority, balancing poor levels of expertise when the corresponding level of authority is high. In the following, we denote by $\boldsymbol{z}_R \in \mathcal{M}_{N,K}$ the latent unified representation encoding both the expertise and the authority of the researchers. In particular, $\boldsymbol{z}_R(i) \in \mathbb{R}^K$ is the expertise vector associated to researcher i and $\boldsymbol{z}_R^k(i) \in \mathbb{R}$ reflects to what extent the researcher have expertise and is authoritative in topic k. Similarly, $\boldsymbol{z}_A \in \mathcal{M}_{M,K}$ is the latent representation encoding the expertise and the authority of the articles.

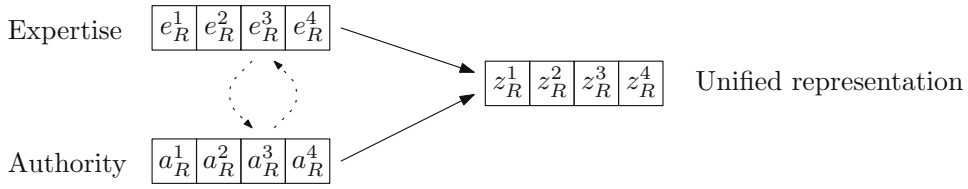

Fig. 3. Illustration of the mutual reinforcement principle between the notions of expertise and authority using a toy example with 4 topics (left) and the proposed unified representation (right).

3.3 Problem Formulation

We capture the cyclic relation between expertise and authority and learn the introduced unified representation by minimizing the objective function $\mathcal{L}_\lambda(\boldsymbol{z}_A, \boldsymbol{z}_W)$ formulated by Eq. (1):

$$\begin{aligned} \mathcal{L}_\lambda(\boldsymbol{z}_A, \boldsymbol{z}_W) \;&=\; \lambda||X_{AW} - \boldsymbol{z}_A \boldsymbol{z}_W^T||_F^2 + (1-\lambda)||X_{AA}\boldsymbol{z}_A - \boldsymbol{z}_A||_F^2 \\ &\text{s.t. } \boldsymbol{z}_W > 0, \boldsymbol{z}_A > 0 \end{aligned} \tag{1}$$

where $\boldsymbol{z}_W \in \mathcal{M}_{W,K}$ is a latent matrix associating to each word a topic distribution, $||.||_F$ is the Frobenius norm, and $\lambda \in [0, 1]$ is a user-parameter that controls the sensitivity of both criteria.

The first part of the objective function $||X_{AW} - \boldsymbol{z}_A \boldsymbol{z}_W^T||_F^2$ corresponds to a standard matrix factorization loss [15] aimed at learning latent topics from the articles content while the second part of the function $||X_{AA}\boldsymbol{z}_A - \boldsymbol{z}_A||_F^2$ is a slight variation of the PageRank formulation [18] applyed on the citation matrix. Note that both parts share the proposed common unified representation $\boldsymbol{z}_A$. In particular, for $\lambda = 1$, a standard non-negative matrix factorization problem is tackled over the article/vocabulary matrix. This standard expertise model, hereafter denoted as NMF, summarizes the articles' content ignoring quality signals. Conversely, for $\lambda = 0$, a PageRank-like algorithm, noted PR, is performed over the citing matrix and only the relative importance of the articles is estimated.

By gathering both objectives around the common variable $\boldsymbol{z}_A$, we force the estimated authority (learned with PR) and expertise (learned with NMF) to influence each other during the optimization. As empirically shown in Sect. 4, authoritative features can help to improve the expertise retrieval phase, and conversely, expertise features can help to identify authoritative researchers. This mutual reinforcement principle between the notions of expertise and authority is the core of our proposed unified approach.

Finding the latent variables associated to the articles is equivalent to solve Eq. (2):

$$(\boldsymbol{z}_A^*, \boldsymbol{z}_W^*) = \underset{\boldsymbol{z}_A, \boldsymbol{z}_W}{\arg\min}\ \mathcal{L}_\lambda(\boldsymbol{z}_A, \boldsymbol{z}_W) \tag{2}$$

Since the Frobenius norm is a convex function, standard gradient descent approaches can be used. In particular, we have:

$$\frac{\partial \mathcal{L}_\lambda}{\partial \boldsymbol{z}_A} = 2\left((1-\lambda)D_{AA}(X_{AA}\boldsymbol{z}_A) - \lambda(X_{AW} - \boldsymbol{z}_A\boldsymbol{z}_W^T)\boldsymbol{z}_W\right)$$
$$\frac{\partial \mathcal{L}_\lambda}{\partial \boldsymbol{z}_W} = -2\lambda(X_{AW} - \boldsymbol{z}_A\boldsymbol{z}_W^T)^T\boldsymbol{z}_A$$

where $D_{AA} = \text{diag}(X_{AA}\mathbb{1} - 1)$, or equivalently, $D_{AA}(i,i) = \sum_{1 \le j \le M} X_{AA}(i,j)$. It should be noted that since the parameters $\boldsymbol{z}_A$ and $\boldsymbol{z}_W$ have KM and KW decision variables respectively, the model $\boldsymbol{\theta} = (\boldsymbol{z}_A, \boldsymbol{z}_W)$ defines a metric space in $\mathbb{R}^{K(M+W)}$. In practice, we solve Eq. (2) using the Limited-Memory BFGS [17] algorithm (L-BFGS), a quasi-Newton method for non-linear optimizations when the number of variables is high (more than 100 million in our case).

Finally, we naturally assimilate a researcher's profile to an aggregation of the obtained latent representation of his/her articles. By summing over the associated articles, we have:

$$\boldsymbol{z}_R^* = X_{RA}\boldsymbol{z}_A^* \tag{3}$$

Therefore, we consider that the researcher r_i is more likely to be an expert in the topic k than the researcher r_j iff $\boldsymbol{z}_R^{*k}(i) > \boldsymbol{z}_R^{*k}(j)$.

4 Experiments

This section is dedicated to the presentation of our results. We evaluate the proposition along two main lines:

1. How well the proposed algorithm can be used to identify authoritative researchers in a digital library. In other words, to what extent expertise features can bring authoritative information.
2. How well the proposed solution can answer to the expert finding task by identifying experts in response to a particular topic query. In other words, to what extent authoritative features can help the expert profiling phase.

We first describe the dataset used for the experiments in Sect. 4.1. Then, Sect. 4.2 presents the protocol for evaluation. Competitors and evaluation metric are introduced in Sects. 4.3 and 4.4 respectively. Finally, quantitative and qualitative results are discussed in Sects. 4.5 and 4.6.

4.1 Data

For the evaluation, 3 data sources were merged to construct several labeled expertise graphs. The Microsoft Academic Search database[8] (MAS), the AMiner platform[9], the Core.edu portal[10], and the induced graphs are detailed thereunder.

Raw data. We made use of the digital library Microsoft Academic Search (MAS) for evaluation. The MAS portal is a semantic network providing a variety of metrics for the research community in addition to literature search. The portal has not been updated since 2013 but is still available online and contains valuable information about roughly 40 million articles and 9 million authors. For the evaluation, all articles and corresponding authors associated to the Computer Science community were crawled. Raw data, including articles titles and abstracts, stored in a relational database represents 4.1 Gb.

Quantitative evaluation. For quantitative evaluation, the AMiner portal was used. The platform provides a public list of $1,270$ experts in the computer science field according to 10 expertise domains from Boosting to Support Vector Machine. From this expert list, roughly 900 experts were retrieved in the MAS dataset to constitute a ground truth. For automatic evaluation purpose, a set of label vectors $\{\mathbf{y}_i\}_{i \leq N}$ with $\mathbf{y}_i \in \mathbb{B}^{10}$ is constructed. In particular, y_i^k is a binary label indicating if the researcher r_i is an expert in the field k ($y_i^k = 1$) or not ($y_i^k = 0$). Note that some researchers are considered as experts in different topics. The 10 considered topics are listed in Table 1.

[8] http://academic.research.microsoft.com/.
[9] https://aminer.org.
[10] http://www.core.edu.au/.

Qualitative evaluation. For qualitative evaluation, we made use of the Core.edu portal. The service provides assessments of major conferences in the Computer Science discipline. Standard labels, from A* for leading venues to C for conferences meeting minimum standards, are used to label the conferences. Specifically, 2,158 conferences published by the Core.edu portal were found in the MAS dataset. We used the associated labels to indirectly measure the articles quality.

Expertise graphs. To evaluate the capacity of the proposal to identify authoritative researchers, 10 expertise graphs were constructed using both previous sources of information. Given a topic k and the associated set of experts, an expertise graph G_k is constructed by iteratively adding in the set of nodes (a) the experts, (b) their co-authors, (c) the associated papers, (d) every citing and cited paper, and (e) each corresponding author. Moreover, to evaluate the capacity of the proposal to identify experts in a particular topic, a complete graph G merging the 10 previously defined expertise graphs is also constructed. Statistics of the different graphs used for the evaluation are summarized in Table 1.

Table 1. Statistics of the 11 graphs used for the experiments.

Expertise graph	Experts	Researchers	Articles
G_0 - Boosting	43	52 228	94 172
G_1 - Data Mining	221	86 786	243 071
G_2 - Information Extraction	72	36 880	80 983
G_3 - Intelligent Agents	28	36 323	60 246
G_4 - Machine Learning	52	37 277	69 025
G_5 - Language Processing	36	20 175	36 684
G_6 - Ontology Alignments	42	30 216	48 601
G_7 - Planning	13	22 809	32 710
G_8 - Semantic Web	274	81 039	244 855
G_9 - Support Vector Machine	70	33 448	60 319
G - All	851	131 303	1 427 317

4.2 Protocol

Preprocessing. The articles' content was processed using the Natural Language Toolkit[11] library for Python. Nouns were extracted from the abstracts and the titles of the articles and those appearing in more than 70% of the articles or in less than 20 articles were removed. From this preprocessing step, a vocabulary of roughly 5 000 words was obtained. Note that we voluntary restrained the size of the vocabulary for efficiency considerations and related sparseness problems.

[11] http://www.nltk.org/.

The remaining words were stemmed using the Lancaster Stemmer. We used TF-IDF weights to model the strength of the relations between words and articles. Thus, $X_{AV}(j, w)$ is the TF-IDF weight of the word w in the article a_j. Finally, to avoid full zero columns in the adjency matrix of the evaluations graphs, every researcher without any authored article and all articles that do not cite any paper were removed.

Optimization. The proposed objective function $\mathcal{L}_\lambda(\boldsymbol{z}_A, \boldsymbol{z}_V)$ was minimized using standard optimization packages for Python[12]. We made used of the Limited-Memory BFGS [17] algorithm (L-BFGS), a quasi-Newton method for non-linear optimizations handling many variables. In practice, the optimization spent roughly three days over the complete graph G. Since L-BFGS approximates the objective function locally and might return local optimums, several optimizations were performed in parallel for each value of λ. Moreover, we made the number k of latent topics vary for each run (from 5 to 100). Only the best runs according to the evaluation metric are presented.

Evaluation. We conducted two series of evaluations. The first one was associated to the authority evaluation while the second one focused on the expertise assessment.

1. We studied the capacity of the proposal to identify authoritative researchers. We wanted to show that considering textual features from articles content may help in identifying authoritative researchers. It should be noted that no reconciliation process between topic query and researchers' profile was performed. To this end, we operated as follows. For each latent topic k, the researchers were ranked by decreasing order of predicted scores $\boldsymbol{z}_R^{*k}$ and the model was evaluated, using the set of labels $\{\mathbf{y}_i^k\}_{i \leq N}$. We report here the best performances over the different discovered latent topics.
2. Secondly, we evaluated the capacity of the models to retrieve experts in response to a particular topic query. The 10 topic queries presented in Table 1 were used for evaluations over the graph G and the set of labels $\{\mathbf{y}_i^k\}_{i \leq N}$ was used as groundtruth. For each topic query q, researchers were ranked according to the vector of scores $\boldsymbol{z}_R^{*k}$ where k corresponds to the latent topic maximizing:

$$k = \arg\max_{k \leq K} \prod_{w \in q} \boldsymbol{z}_W^{*k}(w)$$

where we assumed, for simplicity, that $\boldsymbol{z}_W(w)$ is the entry line in the matrix $\boldsymbol{z}_W$ of the word $w \in \mathcal{W}$.

4.3 Competitors

The proposition, denoted below by **UA** (Unified Approach), was compared to the following state-of-the-art models:

[12] https://www.scipy.org/.

– **PR**. The proposal when $\lambda = 0$. It corresponds to a PageRank-like algorithm capturing the authority of the researchers through the quality of the articles they authored.
– **NMF**. The proposal when $\lambda = 1$. It is a standard non-negative matrix factorization approach capturing latent topics from the article/vocabulary matrix.
– **COS**. A standard Information Retrieval model assuming that the expertise score of a researcher for a topic query q is the cosine similarity between q and a researcher profile. To align with state-of-the-art approaches, a researcher profile was built from a concatenation of the authored articles. Both queries and authors were modeled using bag of words representations and TF-IDF weights.
– **LM**. The model proposed by Balog et al. [1] based on language model formalism. Given a query q, researchers were ranked according to the probability $p(q|\theta_{r_i}) = \prod_{w \in q} p(w|\theta_{r_i})$, where θ_{r_i} encodes the profile of the researcher r_i. In particular, $p(w|\theta_{r_i}) = \sum_{a_j \in \mathcal{A}} p(w|a_j) X_{RA}(i,j)$.
– **LMS**. A smoothed version of the former, also proposed by Balog et al. [1]. Probabilities were smoothed by the frequencies of the corresponding terms in the collection. Formally $\tilde{p}(w|a_j) = \alpha p(w|a_j) + (1-\alpha)p(w|\mathcal{A})$. In our experiments, we set $\alpha = 0.5$.

4.4 Evaluation Metric

The standard classification metric AUC (Area Under the Curve) [4] was used to report the performance of the different classifiers. The metric estimates the probability of ranking a randomly chosen expert higher than a randomly chosen researcher in the final ranking by reporting the area under the ROC curve. Therefore, the closer to 1 the AUC, the better the classifier.

4.5 Quantitative Results

Results for the first set of experiments, associating to the evaluation of the authority, are summarized in Table 2. Results for the expertise assessment are given in Table 3.

Authority evaluation. We discuss here the results associated to the evaluation of the proposed method for identifying authoritative researchers in the different expertise graphs. Interestingly, we observe from Table 2 that for most of the expertise graphs there exists at least one configuration of the proposal that outperforms the PR method. Over the graph G, PR achieves 0.647 while the proposal reaches 0.661 for $\lambda = 0.1$. In general, values of λ around 0.2 improve the baseline of roughly 2%. Intuitively, these results confirm that experts constitute hubs in the different expertise graphs, relatively to the articles (they may write more articles than others) but they also form hubs regarding to the nodes associated to the vocabulary. In other words, the TF-IDF edge weights between nodes associated to articles and words, summarized in the discovered latent topics,

indirectly bring authoritative information. This first important result suggests that representing the textual content of the articles in an expertise graph, in particular by considering words as nodes, can reinforce the discriminative process. It is not surprising that for $\lambda = 1$, although some results are not essentially deceptive, most of them are only slightly better than a random classifier. A single NMF approach, at least over the article/vocabulary matrix, does not suit well for authority modelling.

Table 2. Authority evaluation of the proposal using the AUC metric.

λ	PR	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	NMF
G_0	0.683	0.682	**0.693**	0.681	0.683	0.683	0.678	0.672	0.671	0.663	0.621
G_1	0.666	0.671	0.664	0.666	0.660	**0.672**	0.665	0.671	0.669	0.668	0.659
G_2	0.644	0.643	0.642	**0.653**	0.653	0.641	0.642	0.636	0.639	0.631	0.558
G_3	0.671	**0.684**	0.676	0.675	0.672	0.669	0.681	0.682	0.674	0.662	0.565
G_4	0.674	**0.680**	0.667	0.673	0.675	0.672	0.671	0.677	0.670	0.667	0.582
G_5	0.635	0.636	**0.644**	0.634	0.641	0.629	0.643	0.644	0.637	0.628	0.548
G_6	0.642	0.641	**0.651**	0.639	0.634	0.643	0.648	0.638	0.631	0.622	0.586
G_7	0.688	0.688	**0.694**	0.692	0.691	0.688	0.690	0.688	0.672	0.664	0.612
G_8	**0.667**	0.648	0.655	0.662	0.656	0.647	0.658	0.641	0.654	0.649	0.593
G_9	0.671	**0.674**	0.673	0.672	0.670	0.663	0.665	0.667	0.663	0.666	0.559
G	0.647	**0.661**	0.649	0.653	0.659	0.658	0.656	0.649	0.649	0.651	0.553

Expertise evaluation. Here are discussed the results associated to the expert finding task for the 10 topic queries presented in Table 1 over the graph G. Results associated to the PR method are not available since textual content is not taken into account by this approach and, therefore, matching between query and researchers' profile is not possible. It should be noted that results of the UA method were obtained by minimizing the proposed objective function for $\lambda = 0.2$, $K = 20$ and $|\mathcal{V}| = 5\,300$. Table 3, by reporting the AUC of the five competitors for each topic query, shows that on average, the proposal (UA) outperforms all the competitors (AUC ≈ 0.7), especially the strong baseline LMS (AUC ≈ 0.65). It means that the proposal is more likely to rank the experts higher than the competitors. Considering the queries individually, we observe quite important differences between the performances of LMS and UA. For example, UA is very efficient for retrieving the experts in the Boosting and Planning topic but is outperformed by LMS for the Intelligent Agents or Information Extraction fields. Such irregularities in the results might be explained by the quality of the latent topics and more particularly by the way we have performed the preprocessing step. Topics Boosting and Planning are relatively more specific than others and the associated clusters are easier to learn.

Table 3. Expertise evaluation of the competitors using the AUC metric.

Topic query	NMF	COS	LM	LMS	UA
Boosting	0.829	0.703	0.703	0.703	**0.842**
Data Mining	0.671	0.664	0.635	**0.682**	0.681
Information Extraction	0.607	0.601	0.676	**0.696**	0.623
Intelligent Agents	0.628	0.766	0.676	**0.771**	0.717
Machine Learning	0.745	0.622	0.553	0.635	**0.781**
Language Processing	0.464	0.488	0.492	0.487	**0.567**
Ontology Alignments	0.386	0.492	0.499	0.492	**0.512**
Planning	0.837	0.607	0.617	0.617	**0.904**
Semantic Web	0.541	0.648	0.550	**0.651**	0.622
Support Vector Machine	0.723	0.712	**0.786**	0.779	0.743
Mean	0.643	0.646	0.619	0.651	**0.699**

4.6 Qualitative Results

In this section, we study the publications of the top-5 researchers returned by the different models. The repartition of the conferences classes associated to the publications authored by the different top-5 experts is summarized in Table 4. Results are straightforward. The top-5 experts retrieved by UA publish more than 40% of their articles in A* conferences while this number for the researchers retrieved by other competitors is around 20%. More importantly, only 8% of the articles authored by the experts retrieved by UA are published in C conferences. In general, we see that all competitors that do not integrate any authority feature (i.e., NMF, COS, LM and LMS) lead to similar results in term of class repartition while the proposal is more sensitive to the two extremes. This important result puts forward the interest of considering quality signals for profiling since experts seem to be more concerned by the quality of the productions.

Table 4. Percentage of the publications of the top-5 researchers per conference class.

Model	A*	A	B	C
NMF	20.83	41.66	22.91	14.58
COS	21.90	34.28	28.25	15.55
LM/LMS	21.54	31.64	27.60	19.19
UA	**40.85**	**35.10**	15.74	**8.29**

5 Conclusions

Expert profiling and retrieval constitute challenging problematics for the scientific community. Although authority and expertise are widely studied in literature, these concepts are assumed to be independent biasing expert retrieval to

authors with a large amount of publications. To overcome this issue, we defined a unified model based on an heterogeneous graph representation of digital libraries where authority and expertise vectors are learned simultaneously to capture a mutual reinforcement principle. The evaluation conducted on the Microsoft Academic Search data collection showed that capturing both individuals' expertise and authority significantly outperforms strong baselines. In perspective we will study how to integrate new authoritative criteria such as the co-authoring relation. Temporal and cold-start aspects constitute also challenging questions to refine the results.

References

1. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise Corpora. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2006, pp. 43–50. ACM, New York (2006)
2. Balog, K., de Rijke, M.: Determining expert profiles (with an application to expert finding). In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artifical Intelligence, pp. 2657–2662. Morgan Kaufmann Publishers Inc., San Francisco (2007)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7), 1145–1159 (1997)
5. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2003, pp. 528–531. ACM, New York (2003)
6. Craswell, N., Hawking, D., Vercoustre, A.-M., Wilkins, P.: P@noptic expert: searching for experts not just for documents. In: Ausweb, pp. 21–25 (2001)
7. Davenport, T.H., Prusak, L., Prusak, L.: Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, Boston (1997)
8. de La Robertie, B., Pitarch, Y., Teste, O.: Measuring article quality in Wikipedia using the collaboration network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM 2015, pp. 464–471. ACM, New York (2015)
9. Deng, H., King, I., Lyu, M.R.: Formal models for expert finding on DBLP bibliography data. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM 2008, pp. 163–172. IEEE Computer Society, Washington, D.C. (2008)
10. Gollapalli, S.D., Mitra, P., Giles, C.L.: Ranking experts using author-document-topic graphs. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL 2013, pp. 87–96. ACM, New York (2013)
11. Haveliwala, T.H.: Topic-sensitive pagerank. In: Proceedings of the 11th International Conference on World Wide Web, WWW 2002, pp. 517–526. ACM, New York (2002)
12. Huynh, T., Takasu, A., Masada, T., Hoang, K.: Collaborator recommendation for isolated researchers. In: Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2014, pp. 639–644. IEEE Computer Society, Washington, D.C. (2014)

13. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 919–922. ACM, New York (2007)
14. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
15. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst. **1**, 556–562 (2001)
16. Li, C.-L., Su, Y.-C., Lin, T.-W., Tsai, C.-H., Chang, W.-C., Huang, K.-H., Kuo, T.-M., Lin, S.-W., Lin, Y.-S., Lu, Y.-C., Yang, C.-P., Chang, C.-X., Chin, W.-S., Juan, Y.-C., Tung, H.-Y., Wang, J.-P., Wei, C.-K., Wu, F., Yin, T.-C., Yu, T., Zhuang, Y., Lin, S.-D., Lin, H.-T., Lin, C.-J.: Combination of feature engineering and ranking models for paper-author identification in KDD cup 2013. In: Proceedings of the 2013 KDD Cup 2013 Workshop, KDD Cup 2013, pp. 2:1–2:7. ACM, New York (2013)
17. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**(151), 773–782 (1980)
18. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, pp. 161–172 (1998)
19. Rybak, J., Balog, K., Nørvåg, K.: Temporal expertise profiling. In: Proceedings of the 36th European Conference on Advances in Information Retrieval, ECIR 2014, pp. 540–546 (2014)
20. Serdyukov, P., Taylor, M., Vinay, V., Richardson, M., White, R.W.: Automatic people tagging for expertise profiling in the enterprise. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR 2011 (2011)
21. Tang, J., Yao, L., Zhang, D., Zhang, J.: A combination approach to web user profiling. ACM Trans. Knowl. Discov. Data **5**(1), 2:1–2:44 (2010)
22. Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L., Su, Z.: Topic level expertise search over heterogeneous networks. Mach. Learn. **82**(2), 211–237 (2011)
23. White, S., Smyth, P.: Algorithms for estimating relative importance in networks. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003, pp. 266–275. ACM, New York (2003)
24. Yang, Z., Tang, J., Wang, B., Guo, J., Li, J., Chen, S.: Expert2bole: from expert finding to bole search. In: Knowledge Discovery and Data Mining (2009)
25. Yimam-Seid, D., Kobsa, A.: Expert-finding systems for organizations: problem and domain analysis and the DEMOIR approach. J. Org. Comput. Electron. Commer. **13**(1), 1–24 (2003)
26. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 221–230. ACM, New York (2007)
27. Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B.L., Zha, H., Giles, C.L.: Learning multiple graphs for document recommendations. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, pp. 141–150. ACM, New York (2008)