Identifying Authoritative Researchers in Digital Libraries using External a Priori Knowledge

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

Identifying Authoritative Researchers in Digital Libraries Using External a Priori Knowledge

B. de La Robertie
Université de Toulouse
IRIT UMR5505, F-31071, France
baptiste.delarobertie@irit.fr

Y. Pitarch
Université de Toulouse
IRIT UMR5505, F-31071, France
yoann.pitarch@irit.fr

A. Takasu
National Institute of Informatics
Hitotsunashi, Chiyoda, Tokyo, Japan
takasu@nii.ac.jp

O. Teste
Université de Toulouse
IRIT UMR5505, F-31071, France
teste@irit.fr

ABSTRACT

Numereous digital library projects mine heterogeneous data from different sources to provide expert finding services. However, a variety of models seek experts as simple sources of information and neglect authority signals. In this paper we address the issue of modelling the authority of researchers in academic networks. A model, *RAC*, is proposed that merges several graph representations and incorporate exter-nal knowledge about the authority of some major scientific conferences to improve the identification of authoritative researchers. Based on the provided structural model a biased label propagation algorithm aimed to strenghten the scores calculation of the labelled entities and their neighbors is developped. Both quantitative and qualitative analyses validate the effectiveness of the proposal. Indeed, RAC outperforms state-of-the-art models on a real-world graph containing more than 5 million nodes constructed using *Microsoft Academic Search*, *AMiner* and *Core.edu* databases.

Keywords

Expert Ranking, Graph Analysis, Digital Libraries

1. INTRODUCTION

Digital libraries are rapidly becoming important sources of information. The task of searching for people, and in particularly experts, i.e., individual with "*tacit knowledge*" [2], has received an important interest due to the increasing availability of both relational and organizational data. Platforms such as *Linkedin*[1], *MAS*[2], *DBLP*[3], *CiteSeer*[4] or more recently *AMiner*[5] are well-known examples of applications exploiting both content and structural properties of harvested digital objects to seek individuals' expertise. In the medical domain, the portal *Expertscape*[6] is another well known example of free search engines aimed to mine experts according to some basic geographic features (country, region, city or institution).

In the scientific litterature, expert finding models [1, 15, 16, 18, 20] mainly focus on the reconcilation process between a query and a set of researchers, often associated to *expert profiling* and *expert finding* phases. They extract representations from heterogeneous document collections, as in corporate intranet [5] or email communications [4], and suppose individuals' publications to be representative of their expertise. Then, standard *Information Retrieval* techniques are used to determine the final ranking. Substantial efforts have been made to model individuals' expertise but the way of computing and/or incorporating authoritative/quality signals in such models is often neglected. Such approaches "*mine documents to determine who knows what.*" [4] but do not tackle any question of *credibility*, *quality* or *authority*.

In this paper, we propose to model the *authority* of the researchers in digital libraries. Structural properties concerning the researchers and their publications as well as qualitative information concerning the authority of some major conferences are used to build the model. The proposed heterogeneous graph allows the propagation of different authority signals while capturing a *mutual reinforcement principle* hold between the quality of the different entities. Indeed, as it will be shown in Section 3, authoritative researchers are more likely to publish in high quality conferences, and, con-

http://dx.doi.org/10.1145/3019612.3019809

[1]https://www.linkedin.com/
[2]http://academic.research.microsoft.com/
[3]http://dblp.uni-trier.de/
[4]http://citeseerx.ist.psu.edu
[5]https://aminer.org
[6]http://expertscape.com/

versely, high quality conferences are more likely to publish papers authored by authoritative researchers. In addition, a fix knowledge about the authority of the conferences is handled by the proposed algorithm, leading to a gradual valorization/depreciation of the associated nodes quality during the propagation.

To summarize, our contributions are as follows:

- We propose a structural model based on a heterogenous graph representation that captures a *mutual reinforcement principle* between the conferences' authority and the publications' quality to identify authoritative researchers;
- We design a biased label propagation algorithm handling *external knowledge* about the authority of some conferences;
- Finally, we conduct both quantitative and qualitative analyses on a substantial real world dataset constructing from three different sources, containing more than 4 million articles and 1 million researchers[7].

The rest of paper is organized as follows: Section 2 discusses the related work. Section 3 introduces the motivations. Section 4 formally presents the model and the algorithm. Experiments are discussed in Section 5. Finally, conclusions and future work are drawn in Section 6.

2. RELATED WORK

Substantial efforts have been made to tackle the *expert profiling* and *expert finding* tasks. Using topic models [1, 16, 2], individual's expertise is modeled based on the content of their contributions (see the survey of Balog et al.[2] for further details). However, no explicit authoritative signal is used to discriminate the quality of the articles neither the researchers' authority. Deng et al. [7] tackle the limitation by considering the citation graph as a unique quality signal. This information is integrated in the topic model as a prior probability. However, as motivated in Section 3, this simple quantitative metric as well as other standard indices based on citations counts, i.e., Hirsch index and variants [9, 3, 14], are not discriminative enough to identify authoritative researchers.

Graph-based models, largely based on random walk [20], are often used to identify important nodes in networks and the correlation between centrality and expertise in organizational networks have been extensively studied [12, 4, 17, 11, 21, 23, 13, 8, 6]. The value of the *co-citation graph* has been proven for web pages [12]. The *co-author graph* has been demonstrated to carry out authority signals in Wikipedia [6]. Both *co-author* and *citation* graphs are used to discriminate researchers' importance rating [10]. Moreover, Campbell et al. [4] suggest that graph-based approaches perform better than content-based ones for the experts finding task, motivating our work.

In all these previous works, no external *a priori* knowledge is integrated in the process. Moreover, proposed random walk algorithms are performed over simple graphs, preventing from propagating different authority signals in the network. Zhou et al. [23] also propose to learn from several graphs but they tackle a document recommendation task which is quite different to the proposal. Moreover, they evaluate their solution over two small datasets containing only 400 and 800 researchers respectively.

Unlike state-of-the-art methods, our scalable proposition combines several graphs and make use of external *a priori* knowledge to aggregate and propagate the quality scores in the underlying graph. To the best of our knowledge, we propose the first biased label propagation algorithm used for an expert finding task.

[7]Data are made available for scientific community at http://sac.com

3. MOTIVATIONS

Based on the *Microsoft Academic Search* dataset (details about the data are given in Section 5), we motivate the need of exploiting the four following families of features to compute the authority of a researcher:

- Quality of the articles he/she authored;
- Quality of the articles that cite his/her publications;
- Authority of the conferences where his/her articles are published;
- The authority of his/her collaborators.

To this end, basic statistics of different *Expert Graphs* and *Researcher Graphs* are compared. As detailled in Section 5.1, the experts list provided by the *AMiner* platform was used to build the following graphs. An *Expert Graph* denoted by G_e is constructed as follows:

1. **Seed** (initial set of nodes): 500 random experts;
2. **Collaborators**: every co-author of the researchers in the seed;
3. **Articles**: all articles authored by the added researchers (seed + collaborators);
4. **Citing Articles**: all articles citing the articles authored by the researchers in the seed;
5. **Conferences**: every conference in which all the previous articles have been published.

The set of edges contains the *authoring*, *co-authoring*, *publishing*, and *citing* relations between the different entities (researchers + articles + venues). Edges' weight associated to the co-authoring relation is the number of articles two researchers have co-authored. In a similar way, a *Researcher Graph* denoted by G_r is the graph where the seed is randomly chosen beyond the "non expert" researchers. Averaged statistics of the *authoring*, *citing*, and *co-authoring* relations over 100 different random instances of G_e and G_r are reported in Table 1 and Table 2. Note that venues' label and terminology from the *Core.edu* portal are used.

Firstly, we observe from Table 1 that not only experts publish more than others but also are get used to publish more in top conferences than others (4 097 articles in A* venues). *The more a researcher publishes high quality articles, the more likely he/she is authoritative.*

Secondly, we note that not only papers authored by experts are cited more but are cited more by papers published in top conferences. *The more the researcher's publications are cited by top conferences, the more likely he/she is authoritative.*

		A*	A	B	C
Art.	G_r	136 ± 35	279 ± 53	279 ± 44	274 ± 54
	G_e	**4097** ± 245	3241 ± 173	2354 ± 120	2753 ± 164
Cit.	G_r	36 ± 58	46 ± 61	22 ± 24	21 ± 21
	G_e	**1257** ± 195	985 ± 200	472 ± 87	460 ± 67

Table 1: Number of articles and citations (± standard deviation) per authority class of conferences in G_r and G_e.

Properties	G_r	G_e
Nodes (Researchers)	3 478	18 789
Edges (Co-authoring relations)	6 486	49 195
Average degree	1.86	2.62
Average weighted degree	3.64	**8.52**
Average path length	2.17	5.195
Average diameter	6.12	13.24

Table 2: Statistics of the collaboration subgraphs of G_r and G_e.

Finally, from Table 2, we note that experts not only involve more collaborations than others but are more get used to collaborate. The average weighted degree is twice as much for the experts (8.52) than for other researchers (3.64), indicating that experts have long lasting collaborations.
We do think there is an interest in *aggregating* and *propagating* these different authority signals in the underlying graph in order to indentify authoritative researchers.

4. MODEL

The digital library is modeled using a heterogeneous graph where the researchers, articles and conferences are the sets of nodes, and the *authoring*, *co-authoring* and *citing* relations are the set of edges. As motivated in the previous section, authoritative researchers are linked to many high quality entities. They collaborate with authoritative researchers, write high quality articles and publish in top conferences. This reinforcement principle is captured with a label propagation algorithm that propagate the authority scores, quality scores and biases in the graph. The injected *a priori knowledge* about the authority of the conferences is modeled using an ordered set defining an user preferences over the conferences. During the propagation, biases associated to each conference are updated so that the user preferences over the conferences is respected. Section 4.1 introduces the notations. The RAC model is presented in Section 4.2. Section 4.3 details the associated algorithm.

4.1 Notations

Let $R = \{r_i\}_{1\leq i\leq n}$, $A = \{a_j\}_{1\leq j\leq m}$ and $C = \{c_k\}_{1\leq k\leq p}$ be the sets of n researchers, m articles and p conferences respectively. We suppose that C is a *partially ordered set* modeling a user preferences, i.e., $\exists (c_i, c_j) \in C \times C : c_i \preceq c_j$ or $c_i \succeq c_j$. Let $G = (U, V)$ be a heterogeneous graph defined over the three sets of entities $U = R \cup A \cup C$ and the relations $V = V_{RA} \cup V_{RR} \cup V_{AC} \cup V_{AA}$.

Definition 4.1. *(**Authoring relation** V_{RA}) There exists an edge $(r_i, a_j) \in V_{RA} \subseteq R \times A$ iff researcher r_i is an author of article a_j.*

Definition 4.2. *(**Co-authoring relation** V_{RR}) There exists an edge $(r_i, r_j) \in V_{RR} \subseteq R \times R$ iff researchers r_i and r_j have co-authored at least one article.*

Definition 4.3. *(**Publishing relation** V_{AC}) There exists an edge $(a_i, c_k) \in V_{AC} \subseteq A \times C$ iff article a_i has been published in conference c_k.*

Definition 4.4. *(**Citing relation** V_{AA}) There exists an edge $(a_i, a_j) \in V_{AA} \subseteq A \times A$ iff article a_i cites articles a_j.*

Let M_{RA}, M_{RR}, M_{AC} and M_{AA} be the adjacency matrices associated to the relations V_{RA}, V_{RR}, V_{AC} and V_{AA} respectively. We denote by $\mathbf{q}_R \in \mathbb{R}^n$, $\mathbf{q}_A \in \mathbb{R}^m$ and $\mathbf{q}_C \in \mathbb{R}^p$ the column quality vectors associated to the researchers, articles and conferences respectively. In particular, $\mathbf{q}_C^i$ is the score of conference c_i.

4.2 RAC

We propose to capture the *mutual reinforcement principle* between the quality of the different families of entities in G by solving the following constraint system:

$$\begin{array}{lcl} \mathbf{q}_R & = & M_{RR}\mathbf{q}_R + M_{RA}\mathbf{q}_A \\ \mathbf{q}_A & = & M_{RA}^T\mathbf{q}_R + M_{AA}\mathbf{q}_A + M_{AC}\mathbf{q}_C \\ \mathbf{q}_C & = & M_{AC}^T\mathbf{q}_A + \boldsymbol{\alpha} \\ \text{s.t.} & & \mathbf{q}_C^i \geq \mathbf{q}_C^j \ \forall i,j : c_i \succeq c_j \end{array} \quad (1)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the column bias vector associated to the conferences. Without any prior knowledge ($\boldsymbol{\alpha} = \mathbf{0}$),

- the quality ($\mathbf{q}_R$) of the researchers is defined by the authority of their collaborators ($M_{RR}\mathbf{q}_R$) and the quality of the papers they authored ($M_{RA}\mathbf{q}_A$);
- the quality of an article is function of the authority of its authors ($M_{RA}^T\mathbf{q}_R$), the quality of the papers that cite it ($M_{AA}\mathbf{q}_A$) and the authority of the conference that published it ($M_{AC}\mathbf{q}_C$);
- the quality of a conference is computed using the quality of the published papers ($M_{AC}^T\mathbf{q}_A$).

Intuitively, constraints of Equation (1) force the quality of a conference c_i to be higher than a conference c_j when c_i is prefered over c_j (e.g. c_i is labelled A and c_j is labelled C). The set of constraints implies that for each pair of conferences (c_i, c_j) such that $c_i \succeq c_j$, we should have $\mathbf{q}_C^i - \mathbf{q}_C^j \geq 0$. Finding $\boldsymbol{\alpha}$ satisfiying the introduced constraints is equivalent to minimize the following loss function:

$$\mathcal{L} = \sum_{(i,j):c_i \succeq c_j} \max(0, 1 - (\mathbf{q}_C^i - \mathbf{q}_C^j)) \quad (2)$$

Since $\mathcal{L}$ is convex, a standard gradient descent approach is used to solve $\boldsymbol{\alpha}^* = \arg\min_{\alpha\in\mathbb{R}^p} \mathcal{L}$.

4.3 Algorithm

The proposed label propagation algorithm finds the three fix points $\mathbf{q}_R^*$, $\mathbf{q}_A^*$ and $\mathbf{q}_C^*$, solutions of Equation (1), on the basis of the *power method* [22]. We alternatively compute (a) the vectors of quality scores then (b) the conferences' biases until the scores remain stable. Main steps are the following:

1. Randomly initialize the vectors $\mathbf{q}_R$, $\mathbf{q}_A$, $\mathbf{q}_C$ and $\boldsymbol{\alpha}$;
2. Update the quality scores using Equation (1);
3. Update the bias vectors using Equation (2);
4. Normalize the quality vectors;
5. Check the convergence criterion.

By noting $\mathbf{q}_R^t$, $\mathbf{q}_A^t$, $\mathbf{q}_C^t$ and $\boldsymbol{\alpha}^t$ the vectors computed at the t-th iteration of the algorithm, the propagation stops when the quantity $||\mathbf{q}_R^t - \mathbf{q}_R^{t-1}||_2 + ||\mathbf{q}_A^t - \mathbf{q}_A^{t-1}||_2 + ||\mathbf{q}_C^t - \mathbf{q}_C^{t-1}||_2$ is significantly small, i.e., smaller than $\epsilon \in \mathbb{R}$. More details of the computation are given in Algorithm 1. In our experiments, for $\epsilon = 10^{-6}$, the convergence is reached in less than 10 iterations.

Algorithm 1 Algorithm

Initialize $\mathbf{q}_R^0, \mathbf{q}_A^0$ and $\mathbf{q}_C^0$
Initialize $\boldsymbol{\alpha}^0$
while not converged **do**
  $\mathbf{q}_R^{t+1} \leftarrow M_{RR}\mathbf{q}_R^t + M_{RA}\mathbf{q}_A^t$
  $\mathbf{q}_A^{t+1} \leftarrow M_{RA}^T\mathbf{q}_R^t + M_{AA}\mathbf{q}_A^t + M_{AC}\mathbf{q}_C^t$
  $\mathbf{q}_C^{t+1} \leftarrow M_{AC}^T\mathbf{q}_A^t + \boldsymbol{\alpha}^t$
  $\boldsymbol{\alpha}^{t+1} \leftarrow \arg\min_{\boldsymbol{\alpha}} \mathcal{L}$
  $\mathbf{q}_R^{t+1} \leftarrow \frac{\mathbf{q}_R^{t+1}}{||\mathbf{q}_R^{t+1}||_1}$
  $\mathbf{q}_A^{t+1} \leftarrow \frac{\mathbf{q}_A^{t+1}}{||\mathbf{q}_A^{t+1}||_1}$
  $\mathbf{q}_C^{t+1} \leftarrow \frac{\mathbf{q}_C^{t+1}}{||\mathbf{q}_C^{t+1}||_1}$
end while
return $\mathbf{q}_R^t$

5. EXPERIMENTS

In this section, we first present the data, competitors and metrics used for the experiments. Quantitative and qualitative analysis are performed and discussed in Section 5.5 and Section 5.6 respectively.

5.1 Datasets

Three real-world datasets were used for the experimentations. (1) The *Microsoft Academic Search* (MAS) dataset to construct the graph, (2) the *AMiner* portal to get a ground truth and (3) the *Core.edu*[8] website to get external knowledge about the authority of the conferences.

The *MAS* portal is a semantic network that provides a variety of metrics for the research community in addition to literature search. The service has not been updated since 2013 but remains available and contains valuable information about 39.9 million articles and 9 million authors. Articles and metadata associated to the Computer Science community were crawled to reconstruct the initial graph. Raw data, including abstracts of articles, represents 4.1 Gb. The associated graph contains roughly 5.3 million nodes with 1,190,700 researchers, 4,175,000 articles and 4 700 conferences. In addition, 1,103,000 citations and 8,810,000 coauthorship links are constructed.

The *AMiner* website[9] maintains a list of 1 270 experts in the computer science community, which has already been used to evaluate expert finding models [19]. Among these experts, 1,258 researchers were found in the *MAS* dataset. In this paper, these 1,258 experts are used as a ground truth for evaluation.

Finally, the *Core.edu* portal provides assessments of major conferences in the Computer Science disciplines, from A* for leading venues to C for conferences meeting minimum standards. In particular, $A^* \succ A \succ B \succ C$. From the labeled conferences of the *Core.edu* portal, 2,158 was found in the *MAS* dataset.

[8]http://portal.core.edu.au
[9]https://aminer.org/lab-datasets/expertfinding/

5.2 Competitors

We compare our method to the well established *HITS* [12] algorithm over different authoritative graphs. The following competitors are evaluated:

- **H-ind**. Researchers are ranked by decreasing value of h-index score;
- **H-co**. The *HITS* algorithm is applied over the co-author graph $G_{RR} = (U, V)$ induced by the set of nodes $U = R$ and relations $V = V_{RR}$, and researchers are ranked according to the *Hub* scores;
- **H-cit**. The *HITS* algorithm is applied over the citation graph $G_{AA} = (U, V)$ induced by the set of nodes $U = A$ and relations $V = V_{AA}$. Since it computes an importance score $Hub(a_j)$ for each article a_j, we compute the quality of a researcher r_i as the sum of the articles' scores he authored. Formally $\mathbf{q}_R^j = \sum_{(j,i)\in V_{RA}} Hub(a_i)$;
- **H-rac**. The *HITS* algorithm is applied on the graph G introduced in Section 4.1 and researchers are ranked according to their *Hub* scores;
- **RAC**. The solution as introduced in Section 4.2 without *a priori* knowledge ($\boldsymbol{\alpha} = 0$);
- **RAC+**. The proposed solution incorporating *a priori* knowledge about 2,158 conferences.

5.3 Protocol

Results are averaged over 20 runs with different initializations. For each run, the vector of scores obtained at convergence induces a permutation over the set of researchers. The researchers are ranked by decreasing order of predicted quality. This scheduling is compared to the optimal one consisting in placing the 1,270 true experts provided by the *AMiner* website in top positions.

5.4 Evaluation Metrics

Both classification and ranking metrics are used to evaluate the models. Let $y_i \in \{0, 1\}$ be a flag indicating if researcher r_i is an expert ($y_i = 1$) or not ($y_i = 0$). If σ is the permutation over the researchers induced by the scores at convergence, the *Precision@k* (P@k), *Recall@k* (P@k) and *Normalized Discounted Cumulative Gain* (NDCG@k) are respectively defined by $P@k = \frac{1}{k}\sum_{1\leq i\leq k} y_{\sigma(i)}$, $R@k = \frac{1}{1\,258}\sum_{1\leq i\leq k} y_{\sigma(i)}$ and $DCG(\sigma, k) = \sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}}-1}{\log(1+i)}$ whith $NDCG@k = \frac{DCG(\sigma,k)}{DCG(\sigma^*,k)}$ and σ^* is the optimal ranking over the researchers.

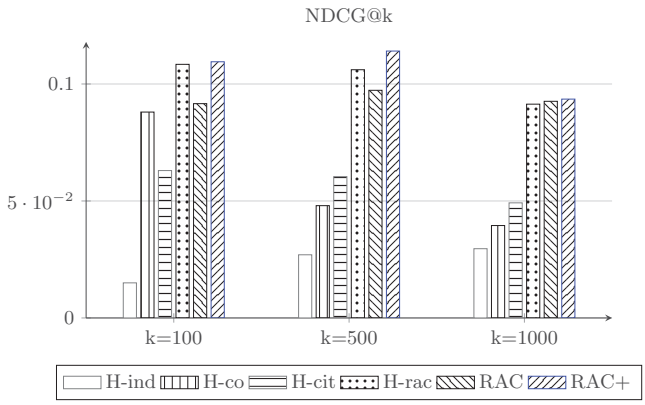

Figure 1: NDCG@k for $k \in \{100, 500, 1000\}$.

5.5 Quantitative Results

Results using the *NDCG* metric are reported in Figure 1. Evaluations using the *Precision* and *Recall* metrics are summarized in Figure 2.

We witness a global increase in the performance of the proposed model compared to classical approaches using different graphs and especially for large values of k. Even without *a priori* knowledge (*RAC* model), the proposal is able to identify more experts than other, confirming the soundness of the heterogeneous representation and in particular the interest of merging different authority patterns in a single representation.

Optimized bias clearly improve the global ranking. Indeed, as illustrated in Figure 1, the *RAC+* model outperforms the competitors. As supported by the qualitative analysis, *RAC+* is able to identify researchers that publish more in top conferences, with authoritative collaborators.

5.6 Qualitative Results

Since the expert list provided by the *AMiner* website can be legitimately discussed, we show the interest of the proposal by studying the publications of the top-5 researchers returned by the different solutions.

From Table 3, we see that both *RAC* and *RAC+* cover more top conferences than other competitors, with 384 and 410 papers published in top conferences respectively. This behaviour is not illustrated by previous quantitative metrics but clearly emphasizes the soundness of our approach.

Table 3: Number of publications produced by the top-5 researchers.

	A*	A	B	C	Total
H-co	377	407	294	**340**	1419
H-ci	290	328	112	86	816
H-rac	381	483	309	248	1421
RAC	384	490	**353**	305	1532
RAC+	**410**	**518**	347	309	**1584**

6. CONCLUSIONS

Digital libraries offer rich heterogeneous relational data to seek expertise. If many works have proposed to identify authoritative individuals by exploiting several different representations, no one has proposed a more general model combining these representations for the expert finding task. In this paper, this task is tackled in the context of academic networks, by merging and exploiting several structural properties of the underlying graph. The proposed model, RAC, exploits a mutual reinforcement principle hold between the authority and the quality of the linked entities. In addition, a lightweight label propagation algorithm is developped, able to handle a priori knowledge about the quality of the conferences to improve the ranking over the researchers. Experiments conducted on the *Microsoft Academic Search* database show the effectiveness of the proposal.

We believe our algorithm can be used to feed classical *Information Retrieval* models and help in identifying experts in some particular domain of expertise. As future work, we will study the interest of the proposal for topics experts finding.

7. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. 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 '06, pages 43–50, New York, NY, USA, 2006.

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

[3] M. Banks. An extension of the hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 2006.

[4] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 528–531, New York, NY, USA, 2003.

[5] N. Craswell, D. Hawking, A. M. Vercoustre, and P. Wilkins. *P@noptic expert: Searching for experts not just for documents*. Ausweb, 2001.

[6] B. de La Robertie, Y. Pitarch, and O. Teste. 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 '15, pages 464–471, New York, NY, USA, 2015.

[7] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 163–172, Washington, DC, USA, 2008. IEEE Computer Society.

[8] S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking experts using author-document-topic graphs. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 87–96, New York, NY, USA, 2013.

[9] J. E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, Dec. 2010.

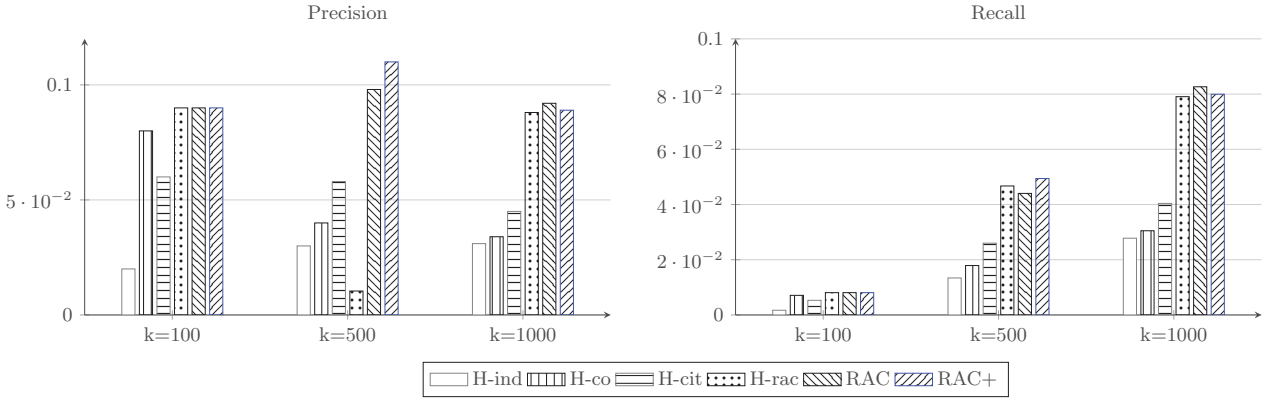

Figure 2: Precision@k and Recall@k for $k \in \{100, 500, 1000\}$.

[10] T. Huynh, A. Takasu, T. Masada, and K. Hoang. Collaborator recommendation for isolated researchers. In *Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops*, WAINA '14, pages 639–644, Washington, DC, USA, 2014. IEEE Computer Society.

[11] P. Jurczyk and E. Agichtein. 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 '07, pages 919–922, New York, NY, USA, 2007.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[13] C.-L. Li, Y.-C. Su, T.-W. Lin, C.-H. Tsai, W.-C. Chang, K.-H. Huang, T.-M. Kuo, S.-W. Lin, Y.-S. Lin, Y.-C. Lu, C.-P. Yang, C.-X. Chang, W.-S. Chin, Y.-C. Juan, H.-Y. Tung, J.-P. Wang, C.-K. Wei, F. Wu, T.-C. Yin, T. Yu, Y. Zhuang, S.-d. Lin, H.-T. Lin, and C.-J. Lin. 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 '13, pages 2:1–2:7, New York, NY, USA, 2013.

[14] C. Moreira, P. Calado, and B. Martins. *Progress in Artificial Intelligence: 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, October 10-13, 2011. Proceedings*, chapter Learning to Rank for Expert Search in Digital Libraries of Academic Publications, pages 431–445. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[15] M. A. Rodriguez and J. Bollen. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 319–328, New York, NY, USA, 2008.

[16] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1133–1142, New York, NY, USA, 2008.

[17] A. Sidiropoulos and Y. Manolopoulos. A citation-based system to assist prize awarding. *SIGMOD Rec.*, 34(4):54–60, Dec. 2005.

[18] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic level expertise search over heterogeneous networks. *Mach. Learn.*, 82(2):211–237, Feb. 2011.

[19] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008.

[20] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 266–275, New York, NY, USA, 2003.

[21] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 221–230, New York, NY, USA, 2007.

[22] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 1036–1043, New York, NY, USA, 2005.

[23] D. Zhou, S. Zhu, K. Yu, X. Song, B. L. Tseng, H. Zha, and C. L. Giles. Learning multiple graphs for document recommendations. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 141–150, New York, NY, USA, 2008.