Expert2Bólè: From Expert Finding to Bólè Search

Zi Yang* Jie Tang* Bo Wang† Jingyi Guo* Juanzi Li* Songcan Chen†
* Dept. of Computer Science, Tsinghua University, China
yangzi@keg.cs.tsinghua.edu.cn, jietang@tsinghua.edu.cn
†Dept. of Computer Science, Nanjing Univ. of Aeronautics and Astronautics, China
bowang@nuaa.edu.cn

ABSTRACT

Expert finding, aiming to answer the question: "Who are experts on topic X?", is becoming one of the biggest challenges for information management. Much work has been conducted for expert finding. Methods based on language model, topic model, and random walk have been proposed. However, little work has studied why people want to find experts.

In this work, we describe Expert2Bólè, a search tool that offers expert finding for various purposes. Specifically, we first employ the *learning-to-rank* techniques to learn a function for ranking experts. We further investigate a specific case of why people search experts, i.e. Bólè search, which tries to identify best supervisors in a given field. How to learn a good ranking function for Bólè search is a very challenging issue, since there would be very limited or even no supervised information which can be used to learn the ranking function. We propose a unified knowledge transfer approach which takes advantage of the expert finding knowledge to learn the ranking function for Bólè search. A prototype system has been developed for expert finding and Bólè search based on the proposed approach. Experiment results show the effectiveness of the proposed approach.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Bólè Search, Expert Finding, Cross Domain Learning

1. INTRODUCTION

With the emergence and rapid proliferation of the social applications and media, search for people has attracted much more attention. One of the most important issues in people search is expert finding. It can be used to answer many challenging questions: how to find the collaborators for a project? how to find an expertise consultant? and how to find a best supervisor? While many methods have been proposed based on, for example, language model[3], topic model[8], and random walk[7], most of existing works primarily focus on the generic expert finding task and little work has studied the problem of expert finding for different purposes. For example, if an undergraduate wants to find a best supervisor with expertise on a topic, is the generic expert finding method sufficient?

We argue that a generic expert finding method is insufficient to find specific expert for various purposes. In this paper, we use Bólè search[1] as a case study to demonstrate why such a consideration is necessary. Bólè search is to find best advisors capable of cultivating novice researchers. This is different from the generic expert finding problem. It is more important for a Bólè to judge and nurture experts than show up his/her own expertise. Mohan et al. propose a propagation algorithm to rank the researchers by their nurture ability[6], but the approach is query-independent and is inapplicable for a search task.

To address these challenges, we propose a general framework, called Expert2Bólè, for expert finding and Bólè search. The key features of Expert2Bólè are the following:

- Expert2Bólè uses a learning-to-rank method to learn a ranking function for expert finding.
- It can identify the common feature space between the generic expert finding and a new specific expert ranking task, which is subsequently used for learning the ranking functions for the new expert finding task.
- An online demonstration system has been developed. Empirical experiments show that the proposed approach performs better (+2.20∼+16.1% in terms of MAP) than the baseline methods.

The proposed framework is quite general and flexible. The learning function and the method for discovering the common latent space can be implemented in many other ways. Variations of the framework can be adapted to many other applications such as collaborator finding and social search.

2. TECHNICAL SPECIFICATION

2.1 Architecture

Expert2Bólè is a framework that aims to support expert finding of both generic and specific purposes. Figure 1 shows

[1]The term Bólè is referred to a Chinese legendary person, who excel in judging and nurturing "thousand-li horses" (a horse that can run a thousand miles without a rest) or "world-class horses".

KDD'09, June 28–July 1, 2009, Paris, France.
ACM 978-1-60558-495-9/09/06.

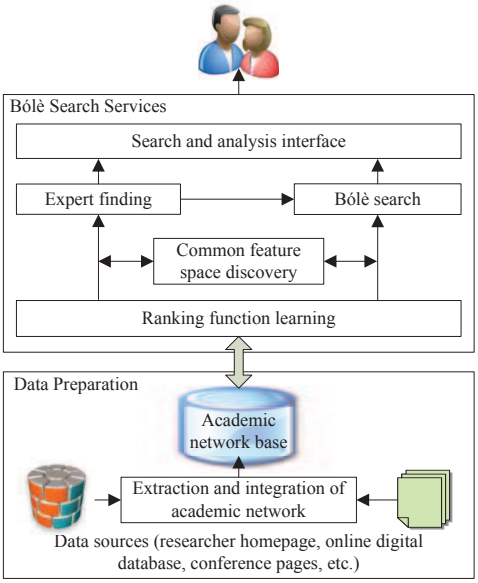

Figure 1: Architecture of Bólè search

the architecture of Expert2Bólè. The system mainly consists of the following components:

1. *Data preparation*: We crawl the academic data from the Web, including researchers' homepages, conference pages, and publications. The extracted data are stored and indexed in an academic network base.
2. *Expert finding*: The generic rank of experts related to a given topic is relatively easy to judge, in terms of how many papers one has published, how many top conference/journal papers one has published, what distinguished awards one has been awarded, etc.[2] With the training dataset, we use the learning-to-rank techniques to learn a function for ranking experts.
3. *Bólè search*: In contrast to generic expert finding, it is difficult to obtain sufficient labeled data for a specific expert finding task such as Bólè search. Thus, in this component, we take advantage of the labeled data for the generic expert finding task. The basic idea is to first discover a common latent feature space between expert finding and Bólè search, and then learn a ranking function for Bólè search by utilizing the supervision information from expert finding via the common feature space.

A demonstration tool has been implemented and is now available at `http://bole.arnetminer.org`. Figure 2 and Figure 3 show two screenshots of the tool. Figure 2 shows three identified Bólès for "machine learning" and Figure 3 shows the example relationships between "Bólès" (denoted by red nodes) and their advisees (denoted by yellow nodes). (The relationships are recognized automatically and the approach will be described in Section 2.4.) The size of each node approximately represents the expertise of the corresponding researcher.

2.2 Data Preparation

For searching and mining the academic network, we need first extract the networking data from the Web. Some of the academic data can be extracted from structured data sources such as the publication information from DBLP; while other information needs to be extracted from unstructured Web pages such as researchers' homepages. We propose a unified approach to extract researcher profiles from their homepages, and integrate the publication data from online databases. The extracted/integrated data are stored into an academic network base. So far, we have already collected 548,504 researcher profiles, 3,378,752 papers, 5,042 conferences, and 32,215,473 citation relationships, 47,443,857 coauthor relationships, and 14,720,130 paper-published-at relationships. Interested readers can refer to [8] for details.

[2] `http://arnetminer.org/lab-datasets/expertfinding/`

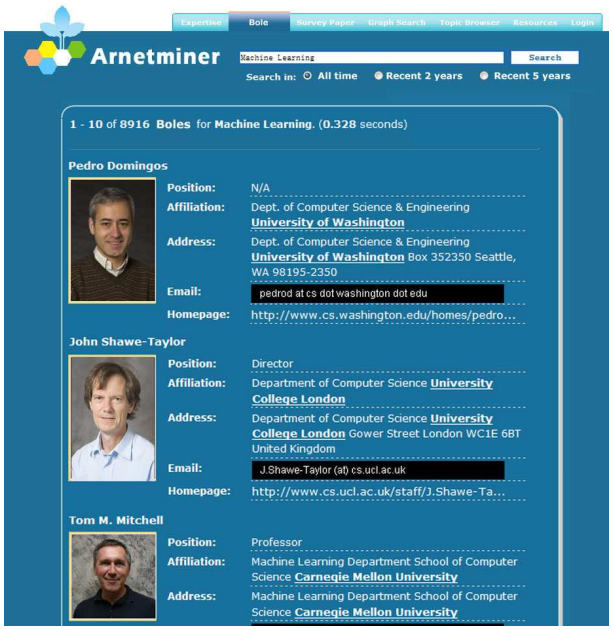

Figure 2: Example of Bólè search for the query "Machine learning".

2.3 Expert Finding

We employ the learning-to-rank techniques for expert finding. As for the learning model, we use Ranking SVM[4], a state-of-the-art supervised rank learning algorithm. Specifically, given a labeled training data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{l}$ and unlabeled test data $\mathcal{S} = \{x'_i\}_{i=1}^{u}$, Ranking SVM aims to learn a ranking function $f \in F$ which can predict the relative order of instances: $x_i \succ x_j \Leftrightarrow f(x_i) > f(x_j)$.

Our labeled training dataset for expert finding including 14,134 persons, 10,716 papers, and 1,434 conferences[3]. There are four-grade scores for each expert: definite expertise, expertise, marginal expertise and no expertise.

For easy explanation, we redefine some notations. Given an instance pair (x_i^a, x_i^b) from different rank levels (y_i^a, y_i^b) within one query, we can create a new instance $(x_i^a - x_i^b, z_i)$ where $z_i = +1$ if $y_i^a > y_i^b$, otherwise $z_i = -1$. Thus, the Ranking SVM mode can be learned from the new training data $\mathcal{L}' = \{(x_i^a - x_i^b, z_i)\}_{i=1}^{n}$ by optimizing:

$$\begin{aligned} &\arg\min_{w^*} \ \ \tfrac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t. \ \ z_i\langle w, x_i^a - x_i^b\rangle \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, \cdots, n \end{aligned} \tag{1}$$

For a new input query, we use $f = \langle w^*, x'\rangle$ to predict the rank level of $x' \in \mathcal{S}$. Experts are then listed in a descending order of their ranking values.

Figure 3: Example relationships between "Bólès" (denoted by red nodes) and their advisees (denoted by yellow nodes).

2.4 Bólè Search

For a specific expert finding task like Bólè search, it is usually difficult to obtain sufficient labeled training data. How can we utilize the existing supervision data (training data) of the generic expert finding for Bólè search? Our basic idea is to find a common feature space in which data for expert finding and Bólè search have similar representation and meanwhile all the partial ranks can be preserved. If we can find a "good" common feature space, the knowledge captured in the common feature space for expert finding can be useful for ranking Bólès.

We propose a unified approach that can simultaneously discover the latent space while learning the ranking function for Bólè search. For expert finding (source domain), we have a training dataset consisting of n_S instance pairs $\mathcal{L}'_S = \{(x^a_{S_i} - x^b_{S_i}, z_{S_i})\}_{i=1}^{n_S}$. For Bólè search (target domain), we assume there are only a few labeled training instances (of size n_T) $\mathcal{L}'_T = \{(x^a_{T_i} - x^b_{T_i}, z_{T_i})\}_{i=1}^{n_T}$.

We denote two ranking functions f_S and f_T respectively for expert finding and Bólè search, which are connected by the latent space U. Therefore, we optimize the objective function by minimizing the two ranking loss functions simultaneously in order to find the best latent space:

$$\min_{W,U} \sum_{t\in\{S,T\}} \sum_{i=1}^{n_t} C_t[1 - z_{t_i}\langle w_t, U^\top(x^a_{t_i} - x^b_{t_i})\rangle]_+ + \lambda||W||^2_{2,1} \quad (2)$$
$$s.t. \quad U^\top U = I$$

where $||W||^2_{2,1}$ is a regularizer which ensures the learned latent space to be common across domains; C_S and C_T are cost-sensitive factors for source and target domain respectively which deal with the problem of imbalanced labeled instances across domains, we can simply set the ratio $\frac{C_T}{C_S}$ to be constant; λ balances the empirical loss and the regularizer; the projection matrix U denotes the latent space; and the orthonormal constraint makes the matrix U unique.

Because the hinge loss of Ranking SVM in the latent space is convex in U, there is an equivalent convex formulation for Eq. (2) [1]:

Table 1: Features for expert search and Bólè search

Feature	Description
L1-L10	Low-level language model features, refer to [5]
H1-H3	High-level language model featuresz, refer to [5]
S1	The year he/she published his/her first paper
S2	The number of papers of an expert
S3	The number of papers in recent 2 years
S4	The number of papers in recent 5 years
S5	The number of citations of all his/her papers
S6	The number of papers cited more than 5 times
S7	The numebr of papers cited more than 10 times
S8	PageRank score, refer to [7]
SumCo1-8	The sum of coauthors' S1-S8 scores
AvgCo1-8	The average of coauthors' S1-S8 scores
SumStu1-8	The sum of his/her advisees' S1-S8 scores
AvgStu1-8	The average of his/her advisees' S1-S8 scores

$$\min_{M,D} \sum_{t\in\{S,T\}} \left(\sum_{i=1}^{n_t} C_t[1 - z_{t_i}\langle \alpha_t, x^a_{t_i} - x^b_{t_i}\rangle]_+ + \lambda\langle \alpha_t, D^+\alpha_t\rangle \right) \quad (3)$$
$$s.t. \quad D \succeq 0, \mathrm{trace}(D) \le 1, \mathrm{range}(M) \subseteq \mathrm{range}(D)$$

where $M = [\alpha_S, \alpha_T] = UW$, $D = U\mathrm{Diag}\left(\frac{||a^i||_2}{||W||_{2,1}}\right)U^\top$ and superscript "+" indicates the pseudoinverse. For a $p \times q$ matrix X, $\mathtt{range}(X) = \{x|Xz = x, \text{for some } z \in R^q\}$. We omitted the details due to space limitation.

Feature Definition

We define 21 common features for expert finding and Bólè search (as shown in Table 1). Features L1-L10 and H1-H3 are scores calculated using language models, while features S1-S8 represent the expertise scores of an author from different aspects. In addition, we define another 32 special features for Bólè search. SumCo1-SumCo8 represent the overall expertise of one's coauthors, and we average SumCo1-SumCo8 scores over the total number of his/her coauthors, denoted by AvgCo1-AvgCo8. Similarly, we consider the summation and average of the expertise of only his/her advisees through features SumStu1-SumStu8 and AvgStu1-AvgStu8. For SumStu1-SumStu8 and AvgStu1-AvgStu8, we need identify the adviser-advisee relationship between researchers. We employ a heuristic-based method for that. Four features (as shown in Table 2) are defined to identify the adviser-advisee relationship. For any two researchers i and j, we calculate a score $s_{ij} = \sum_k \lambda_k f_k(i,j)$, where weight $\{\lambda\}$ of the features is predefined. Finally, if $s_{ij} > r$, we say author i is the advisor of author j; if $s_{ij} < -r$, we say author i is advised by author j, where r is a predefined threshold, and usually takes 2.5∼3.5. Experiments show that the accuracy of relationship identification with this method is 67.0%.

2.5 Experimental Results

In this section, we focus on evaluating the performance of Bólè search. For expert finding, interested readers can refer to [7]. The dataset for Bólè search consists of 9 most frequent queries, and for each query, we chose top ranked 50 researchers by ArnetMiner.org and chose another 50 researchers who start publishing papers only in recent years (>2003, 91.6% of them are currently graduates or postdoctoral researchers). We sent to each of the researchers an email, in which we listed top 50 researchers for each query, and asked them to give feedbacks on whether each candidate is Bólè ("yes") or not ("no"), or "not sure". Participants can also add other Bólès. Based on the feedbacks from the par-

Table 3: Result of Bólè search

Computer vision	Information retrieval	Machine learning	Semantic web	Support vector machine
Thomas S. Huang	W. Bruce Croft	Geoffrey E. Hinton	Jeff Heflin	Bernhard Scholkopf
Alex Pentland	Hector Garcia-Molina	Sanjay Jain	Timothy W. Finin	Vladimir Vapnik
Azriel Rosenfeld	Norbert Fuhr	Michael I. Jordan	Amit P. Sheth	John Shawe-Taylor
Takeo Kanade	Gerard Salton	Tom M. Mitchell	James A. Hendler	Alex J. Smola
Tomaso Poggio	Fabio Crestani	Avrim Blum	Steffen Staab	Thomas Hofmann

Table 2: Features for relationship identification

Feature	Description	Formula
f_1	Coauthor paper ratio	$\frac{n_{\text{co}}}{n_i} - \frac{n_{\text{co}}}{n_j}$
f_2	Absolute paper difference	$g\left(\frac{n_i - n_j}{N}\right)$
f_3	Year of first paper	$g\left(\frac{t_j - t_i}{T}\right)$
f_4	Time interval until cooperation	$g\left(t_{\text{co}} - t_i\right) - g\left(t_{\text{co}} - t_j\right)$

Note: Notation n_i is the number of publications of author i, and n_{co} is the number of cooperation publications, t_i is the year of author i's first publication, and t_{co} is the first year of coauthors' cooperations. Notation N is a constant that describes the average difference of number of publications between an ordinary teacher and a student, and T is the time interval between their first publications. We take $N = 10$ and $T = 10$ in our experiments. $g(x)$ is an identity function if $-1 < x < 1$ and a sign function if $x \leq -1$ or $x \geq 1$.

Table 4: Evaluation of Bólè search with baselines

	p@5	p@10	p@15	MAP	N@5	N@10
Our approach	**.8285**	.7857	**.8571**	**.7971**	**.6189**	**.6112**
RSVM	.7714	**.8429**	.8285	.7756	.5545	.5947
RSVMt	.8000	.8286	.8476	.7837	.5923	.5999
Language model	.6250	.6875	.6500	.6726	.3343	.3809
Expert finding	.5500	.6000	.6333	.6356	.2102	.2454

ticipants, we organized a list for evaluating Bólè search. We rated each candidate person by simply counting the number of "yes"(+1) and "no" (-1) from the received feedback, and averaged the rates over the number of the corresponding definite feedbacks ("yes" and "no"). In this way, we created a relatively commonly accepted Bólè list for each query. Survey letters, candidate lists as well as commonly accepted Bólè lists will be online available. [3]

We use RSVM (Ranking SVM learned from $\mathcal{L}'_T$) and RSVMt (Ranking SVM learned from $\mathcal{L}'_S \cup \mathcal{L}'_T$)[4] as two supervised baseline methods, and use language model[8] as an unsupervised baseline, which calculates the relevance between the query and the author. We evaluate the results in terms of precision, mean average precision (MAP), and normalized discount cumulative gain (NDCG)[2].

In the experiment, we sample two queries to form the labeled training data, and the rest are treated as test data. Table 4 shows the performance of Bólè search. We see that the proposed method performs better than the baseline methods that directly apply RSVM, RSVMt and language model to Bólè search. To demonstrate the motivation of Expert2Bólè, we also compare the result with the one using generic expert finding method for Bólè task. From the result, we can conclude that the method for generic expert finding is inappropriate for Bólè search (-17% than our approach in term of MAP). Table 3 shows five examples results of Bólès search by the proposed approach.

3. DEMONSTRATION PLAN

We will present our Expert2Bólè search tool in the following aspects.

1. First, we will use a poster to give an overview of the tool, including the motivation and major issues addressed in the system. We will introduce the architecture and the main features of the system.
2. Next, we will describe the *learning-to-rank* approach for expert finding. And then, we will briefly introduce our method for social relationship identification, and present the transfer ranking techniques for learning the ranking function for Bólè via the discovered common latent feature space.
3. After that, we will demonstrate the search services, including expert finding and Bólè search service. The audience will gain more detailed understanding about the significance of the specific expert finding and essence of the transfer ranking technique.
4. Finally, we will share our thoughts about the strength and the weakness of the system. We will further discuss the future work of the system.

Please note that this is an ongoing project. Visitors should expect the system to change. For example, some suggest that the Bólè search is useful and it can be enhanced by adding several new features, and some other people expect Expert2Bólè to be more personalized and customized.

4. ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (60703059), National High-tech R&D Program (2009AA01Z138), Chinese National Key Foundation Research (2007CB310803), and Chinese Young Faculty Research Fund (20070003093).

5. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS 2007*, pages 41–48, 2007.
[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR'2006*, pages 43–55, 2006.
[4] R. Herbrich, T. Graepel, and K. Obermayer. *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA, 2000.
[5] T.-Y. Liu, T. Qin, J. Xu, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *LR4IR 2007*, 2007.
[6] B. K. Mohan. The best nurturers in computer science research. In *SDM*, 2005.
[7] J. Tang, R. Jin, and J. Zhang. A topic modeling approach and its integration into the random walk framework for academic search. In *ICDM'08*, 2008.
[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[3] `http://arnetminer.org/lab-datasets/bole/`