# Extraction and Mining of an Academic Social Network

Jie Tang
Department of Computer Science
Tsinghua University
FIT 1-308, Tsinghua University,
Beijing, 100084, China
jietang@tsinghua.edu.cn

Jing Zhang
Department of Computer Science
Tsinghua University
FIT 1-308, Tsinghua University,
Beijing, 100084, China
zhangjing@keg.cs.tsinghua.edu.cn

Limin Yao, Juanzi Li
Department of Computer Science
Tsinghua University
FIT 1-308, Tsinghua University,
Beijing, 100084, China
{ylm, ljz}@keg.cs.tsinghua.edu.cn

## ABSTRACT
This paper addresses several key issues in extraction and mining of an academic social network: 1) extraction of a researcher social network from the existing Web; 2) integration of the publications from existing digital libraries; 3) expertise search on a given topic; and 4) association search between researchers. We developed a social network system, called ArnetMiner, based on proposed methods to the above problems. In total, 448,470 researcher profiles and 981,599 publications were extracted/integrated after the system having been in operation for two years. The paper describes the architecture and main features of the system. It also briefly presents the experimental results of the proposed methods.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval, Digital Libraries, I.2.6 [**Artificial Intelligence**]: Learning, H.2.8 [**Database Management**]: Database Applications.

## General Terms
Algorithms, Experimentation

## Keywords
Social Network, Information Extraction, Expertise Search

## 1. INTRODUCTION
The quickly growing up Web based social network applications provides abundant data for mining, at the same time bring big challenges to the field. In this paper, we present a system called ArnetMiner (http://www.arnetminer.org). Our objective in this system is to provide services for managing academic social networks, specifically including: 1) how to extract researcher profiles from the Web, 2) how to integrate the researcher profiles and publications, 3) how to simultaneously find expertise objects (of different types) on a topic, and 4) how to find associations between researchers.

## 2. ARNETMINER
Figure 1 shows the architecture of the ArnetMiner system. The system mainly consists of five main components:

1. *Extraction*: it focuses on automatically extracting the researcher profile from the Web.
2. *Integration*: it integrates the extracted researcher profiles and crawled publications.
3. *Storage and Access*: it provides storage and indexing for the extracted/integrated data in the RNKB.
4. *Search*: it provides three types of searches: person search, publication search, and category based search.
5. *Mining*: it provides mining services, e.g., expertise search on a given topic and people association finding.

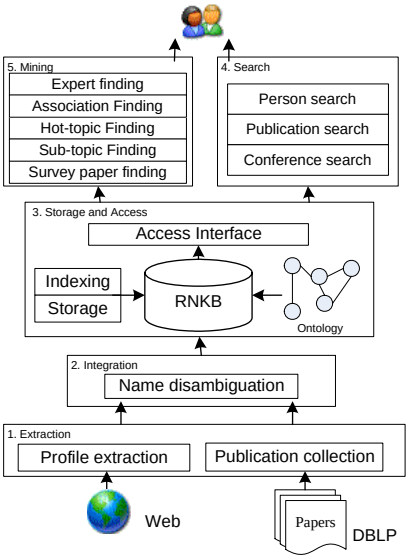


**Figure 1. Architecture of ArnetMiner**

For several features in the system, e.g., extraction of researcher profiles, name disambiguation in the integration, and expertise search, we propose new approaches trying to overcome the drawbacks that exist in the conventional methods. For some other features, e.g., storage and access and searching, we utilize the state-of-the-art methods.

### 2.1 Researcher Profiling
We define the schema of the researcher profile, by extending the FOAF ontology [2]. In the profile, 24 properties and two relations are defined. It is non-trivial to perform the profile extraction, as the layout and content of the researcher homepages/introducing pages may vary largely depending on the authors.

We propose a unified approach to the problem [5]. The approach consists of three steps: relevant page identification, preprocessing and tagging. In relevant page identification, given a researcher name, we first get a list of web pages by a search engine (we used Google API) and then identify the homepage/introducing page using a classifier. The performance of the classifier is 92.39% in terms of F1-measure. In preprocessing, (a) we separate the text into tokens and (b) we assign possible tags to each token. The tokens form the basic units and the pages form the sequences of


*WWW 2008*, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

units in the tagging problem. In tagging, given a sequence of units, we determine the most likely corresponding sequence of tags by using a trained tagging model. (The type of the tags corresponds to the property defined in the profile.) As the tagging model, we use Conditional Random Fields (CRFs) [3]. CRF is a conditional probability of a sequence of labels given a sequence of observations tokens. The CRF is used to find the sequence of tags having the highest likelihood using a trained model. Features were defined for different types of tokens in the CRF model.

For evaluating our unified profiling method, we randomly chose 1,000 researcher names from ArnetMiner and conducted human annotation. Experimental results show that our proposed approach can achieve a performance of 83.37% on average in terms of F1-measure, against Support Vector Machine based method (73.57%) and Amilcare (53.44%).

## 2.2 Name Disambiguation

We integrate the publication data from existing online data source. We chose DBLP bibliography (dblp.uni-trier.de/). For integrating the researcher profiles and the publications, we inevitably have the ambiguous problem. The problem can be described as: Given a person name *a*, let all publications containing the author named *a* as $P=\{p_1, p_2, \ldots, p_n\}$. Suppose there existing *k* actual researchers $\{y_1, y_2, \ldots, y_k\}$ having the name *a*, our task is to assign these *n* publications to their real researcher $y_i$.

Our method is based on a unified probabilistic model based on Hidden Markov Random Fields (HMRF) [6]. This model incorporates constraints and a parameterized-distance measure. The disambiguation problem is cast as assigning a tag to each paper with each tag representing an actual researcher $y_i$.

Specifically, we define the a-posteriori probability as the objective function. We aims at finding the maximum of the objective function. The objective function is defined as the conditional probability of researcher labels *y* given the papers *x*:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left( \sum_{i,k} D(x_i, y_k) + \sum_{i, j \neq i} \{ D(x_i, x_j) \sum_{c_k \in C} [w_k c_k(x_i, x_j)] \} \right)$$

where $D(x_i, y_k)$ is the distance between paper $x_i$ and researcher $y_k$ and $D(x_i, x_j)$ is the distance between paper $x_i$ and $x_j$; $c_k(x_i, x_j)$ denotes a constraint of $x_i$ and $x_j$; $w_j$ is the parameter; and $Z(x)$ is the normalization factor. A Expectation Maximization (EM) based method is used to learn the parameters for the distance function $D(.)$ in the model.

We define six types of constraints based on the characteristic of the publications, e.g., a constraint means two publications have a co-author with the same name. See [6] for the other constraints.

To evaluate our method, we created two datasets, namely Abbreviated Name and Real Name. The first dataset contains 10 abbreviated names (e.g. 'C. Chang') and the second data set has two real person names (e.g. 'Jing Zhang'). The proposed method can obtain an overall performance of 83.0% in terms of pairwise-F1-measure [6], outperforming the baseline [4] by 8.0%.

## 2.3 Expertise Search

The goal of expertise search is aimed at answering: "Who are experts or which are expertise conferences/papers on topic X?".

Traditional, the problem is usually viewed as a ranking problem using either language model to directly calculate the relevance between the query and the object (e.g., paper and author) or random walk model to estimate importance of each object.

We propose a Latent Dirichlet Allocation-style model [1], called Author-Conference-Topic (ACT) model to model the dependencies between different types of objects in the researcher network. In the ACT model, for each paper, an author is first drawn from a uniform distribution; a topic *z* is then drawn from a mixture weight of the chosen author and a distribution from a symmetric Dirichlet prior; next a word is generated from the topic *z* and a conference stamp is generated from the topic *z*. In this way, the dependencies between different types of objects are modeled using the topic *z*. Another advantage of the model is that we can use this model to capture the 'semantic'/hidden relevance between the query and the target objects.

After applying the ACT model to the research network, we again employ a random walk model on the heterogeneous network and finally output a combined score for each object to the query.

We conducted experiments on Arnetminer with seven queries and compared the results with two baselines of using language model and PageRank, as well as results of two existing systems (Libra and Rexa). Experimental results show that the proposed method outperforms them from 4.26% to 29.2% in terms of MAP [7].

## 2.4 Association Search

Finally, we investigate the problem of association search: finding connections between researchers. We formalize the association search as that of near-shortest paths and use a two stage approach to deal with it. First, we employed a shortest path search to find shortest path from all persons in the network to the target person and then we use a depth-first search method to find top *K* ranked results. Our method can find the top *K* results in 2-5 seconds for a general query on the social network with about half million researchers and 1 million publications.

# 3. CONCLUSION

In this paper, we have presented a system called ArnetMiner for extracting and mining a researcher social network. We introduced the architecture and the main features of the system. We have described the four issues that we are focusing on and proposed our approaches to them. Experimental results indicate that the proposed methods can achieve high performances.

# REFERENCES


[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3, 993-1022.

[2] D. Brickley and L. Miller, FOAF vocabulary specification, namespace document, 2004. http://xmlns.com/foaf/0.1/.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. of ICML'2001.

[4] Y.F. Tan, M. Kan, and D. Lee. Search engine driven author disambiguation. Proc. of JCDL'2006. pp. 314-315.

[5] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. Proc. of ICDM'2007. pp. 292-301

[6] D. Zhang, J. Tang, J. Li, and K. Wang. A constraint-based probabilistic framework for name disambiguation. Proc. of CIKM'2007. pp. 1019-1022

[7] J. Zhang, J. Tang, L. Liu, and J. Li. A mixture model for expert finding. Proc. of PAKDD'2008. (to appear)