A Constraint-Based Probabilistic Framework for Name Disambiguation

Duo Zhang, Jie Tang, Juanzi Li, and Kehong Wang
Department of Computer Science and Technology
Tsinghua University, China
{zhangduo, tangjie, ljz, wkh}@keg.cs.tsinghua.edu.cn

ABSTRACT

This paper is concerned with the problem of name disambiguation. By name disambiguation, we mean distinguishing persons with the same name. It is a critical problem in many knowledge management applications. Despite much research work has been conducted, the problem is still not resolved and becomes even more serious, in particular with the popularity of Web 2.0. Previously, name disambiguation was often undertaken in either a supervised or unsupervised fashion. This paper first gives a constraint-based probabilistic model for semi-supervised name disambiguation. Specifically, we focus on investigating the problem in an academic researcher social network (http://arnetminer.org). The framework combines constraints and Euclidean distance learning, and allows the user to refine the disambiguation results. Experimental results on the researcher social network show that the proposed framework significantly outperforms the baseline method using unsupervised hierarchical clustering algorithm.

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: Name Disambiguation, Social Network Analysis, Digital Library, Semi-supervised Clustering

1. INTRODUCTION

The name disambiguation problem can be formalized as partitioning collections of citations into clusters, with each cluster containing only citations authored by the same author, thus disambiguating authorship in citations to induce author name identities [7].

Name disambiguation is a very critical problem in many knowledge management applications, such as Digital Libraries (like Citeseer and DBLP bibliographiy) and Semantic Web applications (like semantic integration and ontology merging). Many knowledge management applications need take name disambiguation as the first step. For example, expert finding, people search, expert profiling, and information integration.

In this paper, we focus on investigating the name disambiguation problem in an academic researcher social network, ArnetMiner (http://arnetminer.org) [10]. Specifically, we aim at assigning papers to the right researchers with the same name. We have examined 30 random person names from ArnetMiner and found that more than 60% of the names have the ambiguity problem.

Traditionally, name disambiguation was often undertaken in either a supervised or unsupervised fashion. In this paper, we intend to propose a general semi-supervised framework to combine the advantages of the supervised and unsupervised methods. Specifically, our contributions are as follows:

- We formalize the name disambiguation problem in a constraint based probabilistic framework. The framework can incorporate any types of domain background knowledge or supervised information (e.g., user interaction) to improve the performances of disambiguation.
- We define six types of constraints and formalize them in the framework. We propose employing EM algorithm to learn different distance metric for different persons.
- We conducted empirical verification of the effectiveness of the proposed framework. Experimental results show that the proposed approach outperforms the hierarchical clustering based disambiguation method.

The rest of the paper is organized as follows. In Section 2, we review related work. In Section 3, we formalize the disambiguation problem. In Section 4, we explain our approach to the problem and in Section 5, we give the experimental results. We conclude the paper in Section 6.

2. RELATED WORK

A number of approaches have been proposed to name disambiguation in different applications.

For example, [3] tries to distinguish web pages to different individuals with the same name. They present two unsupervised frameworks for solving this problem: one is based on link structure of the Web pages and the other uses Agglomerative/Conglomerative Double Clustering method. In [8], name disambiguation is conducted on email data, and the authors use a lazy graph walk method based on the links among emails.

There are also many works focusing on name disambiguation in publication data. For example, Han et al. propose an unsupervised learning approach using K-way spectral clustering method [7].

CIKM'07, November 6-8, 2007, Lisboa, Portugal.

They calculate a Gram matrix for each person name and apply K way spectral clustering algorithm to the Gram matrix to get the result. In [4], the authors solve the problem by constructing a reference graph, and two different measures of relational similarity are experimented. See also [1, 9]. These unsupervised methods use a parameter-fixed distance metric in their clustering algorithm, while parameters of our distance metric can be learned during the disambiguation process.

Two supervised methods are proposed in [6] based on Naïve Bayes and Support Vector Machines. The methods learn a specific model for each author name from the train data and use the model to predict whom a new citation is authored by. However, the method is user-dependent. It is impractical to train thousands of models for all person names in a large digital library.

The other type of related work is semi-supervised clustering, e.g. [2] and [5]. [2] proposes a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields. Their model combines the constraints and distance measures. Compared with [2], we define six kinds of constraints and our method generates the constraints automatically.

3. PROBLEM SETTING

We here give a formal definition of the name disambiguation problem. Given a person name a, we denote publications containing the author name a as $P=\{p_1, p_2, \ldots, p_n\}$. Each publication p_i has six attributes as shown in Table 1. Here, each name $a_i^{(j)}$ has an affiliation $a_i^{(j)}.affiliation$ and an email $a_i^{(j)}.email$. We call the first author name $a_i^{(0)}$, which is actually a, as the *principal author* and the others as *secondary authors*. 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.

Table 1. Attributes of each publication

Attribute	Description
$p_i.title$	title of p_i
$p_i.conference$	published conference/journal of p_i
$p_i.year$	published year of p_i
$p_i.abstract$	abstract of p_i
$p_i.authors$	authors name set of p_i $\{a_i^{(0)},a_i^{(1)},\ldots a_i^{(u)}\}$
$pi.references$	reference set of p_i which is denoted as REF_i

Next, we define a set of constraint functions $C=\{c_1, c_2, \ldots, c_m\}$. The constraint function is a Boolean-valued function and is defined as follow:

$$c_l(p_i, p_j) = \begin{cases} 1 & \text{if } p_i \text{ and } p_j \text{ satisfy the constraint } c_l \\ 0 & \text{otherwise} \end{cases}$$

We extracted the attribute values of each paper from several digital libraries, e.g., IEEE, Springer, and ACM by using heuristics.

4. OUR APPROACH

Our method is based on a unified probabilistic model using Hidden Markov Random Fields (HMRF). 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. Six types of constraints are incorporated into the objective function, where constraints are considered as a form of supervision. If one paper's label assignment violates a constraint, it will be penalized in some sense, which in turn affects the result.

4.1 Formalization using HMRF

A HMRF based semi-supervised framework is first introduced by [2]. Based on HMRF, the conditional distribution of the researcher labels y given the observation x (paper) has

$$\max p(Y \mid X) \propto \min f_{obj} = \sum_i \sum_j \{D(x_i, x_j) I(l_i \neq l_j) \sum_{c_k \in C} [w_k c_k(p_i, p_j)]\} + \sum_{x_i \in X} D(x_i, y_{l_i}) + \log Z$$

where l_i is the tag for x_i, $D(x_i, y_{l_i})$ is the distance between paper x_i and researcher y_{l_i} (represented by a set of assigned papers), $D(x_i, x_j)$ is the distance between paper x_i and x_j, w_k is the weight for c_k, and Z is the normalization factor. The key issue here is how to define *constraints* and how to learn the *distance function* for effectively performing the disambiguation task.

4.2 Constraint Selection

We define six types of constraints based on the characteristic of publication dataset as listed in Table 2.

Table 2. Constraints defined in our approach

C	W	Constraint Name	Description
c_1	w_1	CoOrg	$a_i^{(0)}.affiliation = a_j^{(0)}.affiliation$
c_2	w_2	CoAuthor	$\exists\, r, s>0,\ a_i^{(r)}=a_j^{(s)}$
c_3	w_3	Citation	p_i cites p_j or p_j cites p_i
c_4	w_4	CoEmail	$a_i^{(0)}.email = a_j^{(0)}.email$
c_5	w_5	Feedback	Constraints from user feedback
c_6	w_6	τ-CoAuthor	one common author in τ extension

All these constraints are defined between two papers p_i and p_j. Constraint c_1 means the principal authors of two papers are from the same organization. Constraint c_2 means two papers have a secondary author with the same name, and constraint c_3 means a paper cites another paper. Constraint c_4 means the principal authors of two publications have the same email address (this is a strong constraint). Constraint c_5 denotes the user feedback.

We use an example to explain constraint c_6. Suppose p_i has authors 'David Mitchell' and 'Andrew Mark', and p_j has authors 'David Mitchell' and 'Fernando Mulford'. If 'Andrew Mark' and 'Fernando Mulford' also coauthor one publication, then we say p_i and p_j have a 2-CoAuthor constraint. We construct a matrix M (cf. Figure 1) to test whether two papers have a τ-CoAuthor constraint.

In matrix M, $p_1,p_2,\ldots,p_n$ are publications with an author named a. $a_1,a_2,\ldots,a_p$ is the union set of all $p_i.authors$, $i=1,2,\ldots n$, i.e.

$$\{a_1, a_2, \ldots, a_p\} = \bigcup_{i=1}^{n} p_i.authors = \bigcup_{i=1}^{n} \{a_i^{(1)}, a_i^{(2)}, \ldots, a_i^{(u_i)}\}$$

Note that $a_1,a_2,\ldots,a_p$ does not include $a_i^{(0)}$. The sub matrix M_p indicates the relationship between $p_1,p_2,\ldots,p_n$ and initially it is an identity matrix. In sub matrix M_{pa}, an element on row p_i and column a_j is equal to 1 if and only if $a_j \in p_i.authors$, otherwise 0. The matrix M_{ap} is symmetric to M_{pa}. Sub matrix M_a indicates the co-

authorship among $a_1,a_2,\ldots,a_p$. The element on row x_i and column x_j is equal to 1 if and only if a_i and a_j coauthor one publication in our database (not just limited to $p_1,p_2,\ldots,p_n$), otherwise 0.

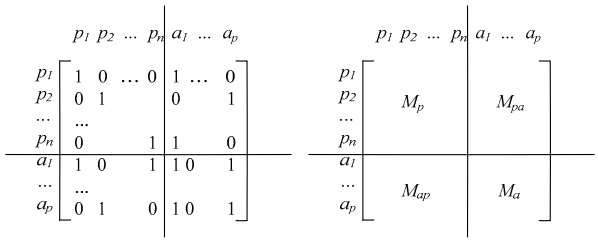

Figure 1. Matrix M for c_6 constraint

By multiplying M by itself, i.e. $M^{(1)}=M \bullet M$, the element on row p_i and column p_j becomes 1 if they have at least one common secondary author. Thus, M shows 1-CoAuthor constraints between papers. Similarly, $M^{(2)}=M^{(1)} \bullet M$ indicates 2-CoAuthor constraints between papers. Likewise for the k-CoAuthor constraints. If p_i and p_j have both τ_1-CoAuthor and τ_2-CoAuthor ($\tau_1<\tau_2$) constraint, we only consider the τ_1-CoAuthor constraint.

Next, we set the weight for each type of constraint empirically. For example, we assign c_2 constraint Co-Author a relatively high weight and assign w_6 as the τ power of w_2, i.e. $w_6= w_2^{\tau}$. Emails can be regarded as unique identifiers for people, so we assign w_4 the largest weight. The larger the weight, the greater the impact of that constraint is. In our experiment, we set $w_1 \sim w_6$ as 0.5, 0.7, 0.6, 1.0, 0.9, 0.7^{τ} respectively.

4.3 EM Framework

Three tasks are executed by the Expectation Maximization (EM) method: learning parameters of the distance measure, re-assignment of paper to researchers, and update of researcher representatives y_k.

We define our distance function $D(x_i, x_j)$ as follows:

$$D(x_i,x_j)=1-\frac{x_i^T \mathbf{A} x_j}{\| x_i \|_{\mathbf{A}} \| x_j \|_{\mathbf{A}}}, \text{ where } \| x_i \|_{\mathbf{A}}=\sqrt{x_i^T \mathbf{A} x_j}$$

here **A** is a parameter matrix. For simplification, we define it as a diagonal matrix.

The EM process can be summarized as follows: in the E-step, given the current researcher representatives (the set of assigned papers), every paper is re-assigned to the researcher by maximizing $p(Y|X)$. In the M-step, the researcher representative y_h is re-estimated from the assignments to maximize $p(Y|X)$ again, and the distance measure is updated to increase the objective function.

For initialization of our EM framework, we first cluster publications into disjoint groups based on the constraints over them, i.e. if two publications have a constraint, then they are assigned to the same researcher. Therefore, we first get λ groups. If λ is equal to the actual researcher number k, then these λ groups are used as our initial assignment. If $\lambda< k$, we choose another (k-λ) clusters by random perturbations of the global centroid. If $\lambda>k$, we cluster the nearest group until there are only k groups left.

In the E-step, assignments of data points to researchers are updated to maximize the $p(Y|X)$. A greedy algorithm is used to sequentially update the assignment of each paper. The algorithm performs assignments in random order for all papers. Each paper x_i is assigned to y_h that minimizes the function:

$$f(y_h,x_i)=\sum_i D(x_i,y_h)+\sum_{i,j\neq i}\{D(x_i,x_j)\sum_{c_k\in C}[w_k c_k(x_i,x_j)]\}$$

The assignment of a paper is performed while keeping assignments of the other papers fixed. The assignment process is repeated after all papers are assigned. This process runs until no paper changes its assignment between two successive iterations.

In the M-step, each researcher representative is updated by the arithmetic mean of its points:

$$y_h=\frac{\sum_{i:l_i=h} x_i}{\| \sum_{i:l_i=h} x_i \|_{\mathbf{A}}}$$

Then, each parameter a_{mm} in **A** is updated by (only parameters on the diagonal): $a_{mm}=a_{mm}+\eta\frac{\partial f_{obj}}{\partial a_{mm}}$, where:

$$\frac{\partial f_{obj}}{\partial a_{mm}}=\sum_{o_i\in O}\frac{\partial f_{obj}}{\partial a_{mm}}+\sum_{i,j\neq i}\{\frac{\partial D(x_i,x_j)}{\partial a_{mm}}\sum_{c_k\in C}[w_k c_k(p_i,p_j)]\}$$

$$\frac{\partial D(x_i,x_j)}{\partial a_{mm}}=\frac{x_i^{(m)}x_j^{(m)}\| x_i \|_{\mathbf{A}}\| x_j \|_{\mathbf{A}}-x_i^T\mathbf{A}x_j\frac{(x_i^{(m)})^2\| x_i \|_{\mathbf{A}}^2+(x_j^{(m)})^2\| x_j \|_{\mathbf{A}}^2}{2\| x_i \|_{\mathbf{A}}\| x_j \|_{\mathbf{A}}}}{\| x_i \|_{\mathbf{A}}^2\| x_j \|_{\mathbf{A}}^2}$$

5. EXPERIMENTS

5.1 Datasets

To evaluate our proposed methods, we created two datasets: Abbreviated Name dataset and Real Name dataset. The first dataset was collected by querying 10 abbreviated names in our database. All the abbreviated names are created by simplifying the original names to its first name initial and last name, for example, 'Cheng Chang' to 'C. Chang'. The simplification form is popular in bibliographic records. Statistics of this dataset is shown in Table 3.

Another dataset is constructed by querying two person names 'Jing Zhang' and 'Yi Li'. The purpose of constructing this dataset is to analyze contributions of the six types of constraints. 'Jing Zhang' has 54 publications by 25 different researchers and 'Yi Li' has 42 publications by 22 different researchers.

Table 3. Abbreviate Name dataset

Abbr. Name	#Publications	#Actual Researcher	Abbr. Name	#Publications	#Actual Researcher
C. Chang	402	97	M. Hong	108	30
G. Wu	152	46	X. Xie	136	36
K. Zhang	293	40	P. Xu	39	5
J. Li	551	102	H. Xu	182	60
B. Liang	55	14	W. Yang	263	82

5.2 Experiment Design

We designed two experiments to evaluate our proposed name disambiguation method.

The first experiment tests the effectiveness of our method with all kinds of constraints included. The experiment was conducted on

the Abbreviate Name dataset. We defined a baseline method based on the previous work. The baseline method uses a hierarchical clustering algorithm to group the papers into clusters. Then we view the grouped papers as the disambiguation results. In the clustering, we used words as features and employed Cosine Similarity Measure to calculate the similarity between two papers. We also suppose that the number of persons k is provided empirically.

To test how constraints of different types affect the performance of disambiguation, we also conducted experiments using different combination of the constraints and compared the results. This experiment was conducted on the Real Name dataset.

In the experiments, we conducted evaluations in terms of pairwise_precision, pairwise_recall, and pairwise_F1-measure [2].

5.3 Name Disambiguation Experiments

5.3.1 Results

The performances of both our method and the baseline methods on the Abbreviation Name dataset are shown in Table 4.

Table 4. Results on Abbreviate Name dataset

	Baseline			Constraint based approach		
Name	Prec.	Rec.	F1	Prec.	Rec.	F1
C. Chang	0.65	0.59	0.62	0.73	0.67	0.70
G. Wu	0.71	0.62	0.66	0.75	0.75	0.75
K. Zhang	0.75	0.60	0.67	0.79	0.71	0.75
J. Li	0.62	0.52	0.57	0.66	0.59	0.62
B. Liang	0.82	0.76	0.79	0.85	0.89	0.87
M. Hong	0.79	0.65	0.71	0.82	0.75	0.78
X. Xie	0.77	0.73	0.75	0.83	0.82	0.82
P. Xu	0.89	0.95	0.92	0.94	1.00	0.97
H. Xu	0.65	0.59	0.62	0.73	0.67	0.70
W. Yang	0.71	0.62	0.66	0.75	0.75	0.75
Avg.	0.75	0.60	0.67	0.79	0.71	0.75

The proposed method outperforms the baseline method by 8.0% in terms of pairwise_F1-measure.

5.3.2 Contribution of Constraints

We investigated the contribution of each type of constraints in name disambiguation. Figure 2 shows the F1 scores of "Jing Zhang" and "Yi Li" in the Real Name dataset with various combinations of constraints. We can see that the CoAuthor constraint contributes a lot to the overall results. It can be also seen that all the constraints we defined can enhance the final performance.

6. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the problem of name disambiguation in social networks. We have proposed a constrain-based probabilistic framework to the problem. The framework can incorporate any kinds of domain background knowledge, as well as user interactions to improve the disambiguation results. We have defined six types of constraints for disambiguating researchers in a real-world social network. We have employed EM algorithm to learn the parameters of the distance measure. Experimental results show that the proposed model significantly outperforms the unsupervised method using hierarchical clustering algorithm.

As future work, we plan to investigate how to estimate the actual researchers number k for a given name automatically, which will greatly enhance the practice of the proposed approach.

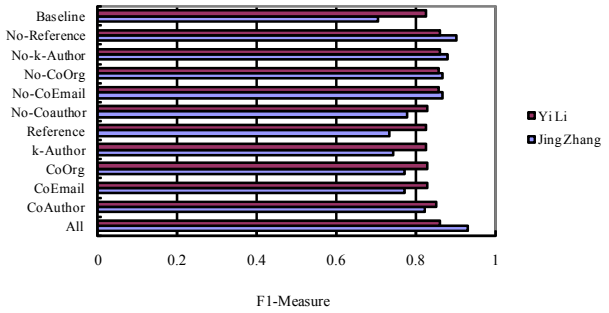

Figure 2. Contribution of constraints

7. ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China under Grant No. 90604025.

8. REFERENCES

[1] Aswani, N., Bontcheva, K., and Cunningham, H. Mining information for instance unification. In *Proceedings of ISWC'2006*, pp.329-342, 2006.

[2] Basu, S., Bilenko, M., and Mooney, R. J. A probabilistic framework for semi-supervised clustering. In *Proceedings of SIGKDD'2004*, pp. 59-68, Seattle, USA, August 2004.

[3] Bekkerman, R. and McCallum, A. Disambiguating web appearances of people in a social network, In *Proceedings of WWW'2005*, pp. 463-470, ACM Press, 2005

[4] Bhattacharya, I. and Getoor, L. Entity resolution in graphs. Book Chapter in Mining Graph Data, Lawrence B. Holder and Diane J. Cook, Editors, Wiley, 2006.

[5] Cohn, D., Caruana, R., and McCallum, A. Semi-supervised clustering with user feedback. *Technical Report* TR2003-1892, Cornell University, 2003

[6] Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of JCDL'2004,* Arizona, USA, pp. 296-305, 2004.

[7] Han, H., Zha, H., and Giles, C. L. Name disambiguation in author citations using a K-way Spectral Clustering Method. In *Proceedings of JCDL'2005,* Denver, Colorado, USA, June 2005, 334 – 343

[8] Minkov, E., Cohen, W. W., and Ng, A. Y. Contextual search and name disambiguation in email using graphs. In *Proceedings of SIGIR'2006*, Washington, USA, pp. 27-34, 2006.

[9] Tan, Y. F., Kan, M., and Lee, D. Search engine driven author disambiguation. In *Proceedings of JCDL'2006*, NC, USA, pp. 314-315, June 2006.

[10] Tang, J., Hong, M., Zhang, J., Liang, B., Yao, L., and Li, J. ArnetMiner: toward building and mining social networks. (Demo). In *Proceedings of SIGKDD'2007*.