A Unified Framework for Name Disambiguation

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

Duo Zhang
Department of Computer Science
University of Illinois at Urbana Champaign
dolphin.zduo@gmail.com

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

ABSTRACT
Name ambiguity problem has been a challenging issue for a long history. In this paper, we intend to make a thorough investigation of the whole problem. Specifically, we formalize the name disambiguation problem in a unified framework. The framework can incorporate both attribute and relationship into a probabilistic model. We explore a dynamic approach for automatically estimating the person number *K* and employ an adaptive distance measure to estimate the distance between objects. Experimental results show that our proposed framework can significantly outperform the baseline method.

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, Probabilistic Model, Digital Library

1. INTRODUCTION
Name disambiguation is a very critical problem in many knowledge management applications, such as Digital Libraries and Semantic Web applications. We have examined one hundred person names and found that the problem is very serious. For example, there are 54 papers authored by 25 "Jing Zhang". Even, there are three "Yi Li" graduated from the author's lab.

In this paper, we propose a unified probabilistic framework to offer solutions to the above challenges. We explore a dynamic approach for estimating the person number *K*. We propose a unified probabilistic model. The model can achieve better performance in name disambiguation than the existing methods.

2. NAME DISAMBIGUATION

2.1 Notations
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.

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

We define five types of undirected relationships between papers. Table 1 shows the relationships. Relationship r_1 represents that two papers are published at the same conference/journal. Relationship r_2 means two papers have a secondary author with the same name, and relationship r_3 means one paper cites the other paper.

Table 1. Relationships between papers

R	*W*	Relation Name	Description
r_1	w_1	Co-Conference	$p_i.pubvenue = p_j.pubvenue$
r_2	w_2	Co-Author	$\exists r, s>0, a_i^{(r)}=a_j^{(s)}$
r_3	w_3	Citation	p_i cites p_j or p_j cites p_i
r_4	w_4	Constraints	Feedbacks supplied by users
r_5	w_5	τ-CoAuthor	τ-extension co-authorship (τ>1)

We use an example to explain relationship r_5. Suppose p_i has authors "David Mitchell" and "Andrew Mark", and p_j has authors "David Mitchell" and "Fernando Mulford". (We are going to disambiguate "David Mitchell".) If "Andrew Mark" and "Fernando Mulford" also coauthor one paper, then we say p_i and p_j have a 2-CoAuthor relationship.

2.2 Our Approach
The proposed framework is based on Hidden Markov Random Fields [1], which can be used to model dependencies between observations (here each paper can be viewed as an observation). Then for each disambiguation task, we propose using the Bayesian Information Criterion (BIC) [2] as the criterion to estimate the person number *K*. We define an objective function for the disambiguation task. Our goal is to optimize a parameter setting that maximizes the objective function with some given *K*.

Figure 1 shows the graphical structure of the HMRF model to the name disambiguation problem. The edge between the hidden variables corresponds to the relationships between papers. The value of each hidden variable indicates the assignment results.

By the fundamental theorem of random fields [3], the probability distribution of the label configuration *Y* has the form:

$$P(Y)=\frac{1}{Z_1}\exp(-\sum_i \sum_{(y_i,y_j)\in E_i} f(y_i,y_j)) \quad (1)$$

where potential function $f(y_i, y_j)$ is a non-negative function defined based on the edge (y_i, y_j) and E_i represents all neighborhoods related to y_i. Z_1 is a normalization factor.

Our goal is to find the maximum a-posteriori (MAP) configuration of the HMRF, i.e. maximize $P(Y|X)$. Suppose $P(X)$ is a constant, then we get $P(Y \mid X) \propto P(Y)P(X \mid Y)$ by the Bayes rule. Therefore, our objective function is defined as:

$$L_{\max} = \log(P(Y \mid X)) = \log(P(Y)P(X \mid Y)) \quad (2)$$

By integrating (1) into (2), we obtain:

$$L_{\max} = \log\left(\frac{1}{Z_1}\exp(-\sum_i \sum_{(y_i,y_j)\in E_i} f(y_i,y_j)) \cdot \prod_{x_i \in X} P(x_i \mid y_i)\right) \quad (3)$$

Then the problem is how to define the potential function f and how to estimate the generative probability $P(x_i|y_i)$.

We define the potential function $f(y_i, y_j)$ by the distance $D(x_i, x_j)$ between paper p_i and p_j. As for the probability $P(x_i|y_i)$, we assume that publications is generated under the spherical Gaussian distribution and thus we have:

$$P(x_i \mid y_i) = \frac{1}{Z_2}\exp(-D(x_i, \mu_{(i)})) \quad (4)$$

where $\mu_{(i)}$ is the cluster centroid that the paper x_i is assigned. Notation $D(x_i, \mu_{(i)})$ represents the distance between paper p_i and its assigned cluster center $\mu_{(i)}$. Thus putting equation (4) into (3), we obtain the objective function in minimizing form:

$$L_{\min} = \sum_i \sum_{(y_i,y_j)\in E_i} f(y_i,y_j) + \sum_{x_i \in X} D(x_i, \mu_{(i)}) + \log Z \quad (5)$$

where $Z=Z_1Z_2$.

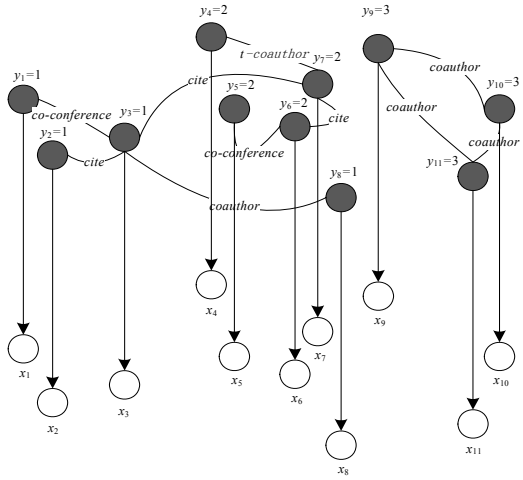

Figure 1. The graphical structure of a HMRF

2.3 EM Algorithm

Three tasks are executed by Expectation Maximization (EM): learning parameters of the distance measure, re-assignment of each paper to a cluster, and update of cluster centroid $u_{(i)}$.

We define the distance function $D(x_i, x_j)$ as follows [1]:

$$D(x_i, x_j) = 1 - \frac{x_i^T \mathbf{A} x_j}{\| x_i \|_{\mathbf{A}} \| x_j \|_{\mathbf{A}}}, \text{ where } \| x_i \|_{\mathbf{A}} = \sqrt{x_i^T \mathbf{A} x_i} \quad (6)$$

here **A** is a parameter matrix.

The EM process can be summarized as follows: in the E-step, given the current cluster centroid, every paper x_i is re-assigned to the cluster by maximizing $p(y_i|x_i)$. In the M-step, the cluster centers are re-estimated based on the assignments to minimize the objective function L; and the parameter matrix is updated to increase the objective function.

2.4 Estimation of *K*

Our proposed strategy (see Algorithm 1) is to start by setting K as 1 and use the BIC score to measure whether to split the current cluster. The algorithm runs iteratively. In each iteration, we try to split every cluster C into two sub-clusters. We calculate a local BIC score of the new sub model M_2. Given BIC(M_2) > BIC(M_1), we split the cluster. We calculate a global BIC score for the obtained new model. The process continues if there exists split. Finally, we use the global BIC score as the criterion to choose as output the model with the highest score. BIC score is defined as

$$BIC^v(M_h) = \log\left(P(M_h \mid P)\right) - \frac{|\lambda|}{2} \cdot \log(n) \quad (7)$$

For the parameter $|\lambda|$, we simply define it as the sum of the K cluster probabilities, weight of the relations, and cluster centroids.

2.5 Experimental Results

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').

We evaluated the performances of our method and the baseline methods (K-means) on the two data sets. Results show that our method can significantly outperform the baseline method for name disambiguation (+46.54% on Abbreviate Name data set and +41.35% on Real Name data set in terms of the average F1-score).

We evaluated the effectiveness of estimation of the person number K. We have found that the estimated numbers by our approach are close to the results by human labeled. We applied X-means to find the person number K. We found that X-means fails to find the actual number. It always outputs only one cluster except "Yi Li" with 2. See [4][5] for more evaluation results.

To further evaluate the effectiveness of our method, we applied it to expert finding and people association finding. For expert finding, in terms of mean average precision (MAP), 2% improvements can be obtained. For people association finding, we selected five pairs of person names and searched in ArnetMiner system, averagely 20% improvements can be obtained.

3. CONCLUSION

In this paper, we have investigated the problem of name disambiguation. We have proposed a generalized probabilistic model to the problem. We have explored a dynamic approach for estimating the person number K. Experiments show that the proposed method significantly outperforms the baseline methods.

4. REFERENCES

[1] S. Basu, M. Bilenko, and R. J. Mooney. A Probabilistic Framework for Semi-Supervised Clustering. In Proc. of SIGKDD'2004, pp. 59-68, Seattle, USA, August 2004

[2] M. Ester, R. Ge, B.J. Gao, Z. Hu, and B. Ben-Moshe. Joint Cluster Analysis of Attribute Data and Relationship Data: the Connected K-center Problem. In Proc. of SDM'2006.

[3] J. Hammersley and P. Clifford. Markov Fields on Finite Graphs and Lattices. Unpublished manuscript. 1971.

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

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