Generative Models for Name Disambiguation

Yang Song[1], Jian Huang[2], Isaac G. Councill[2], Jia Li[3,1], C. Lee Giles[2,1]

[1]Department of Computer Science and Engineering,[2]Information Sciences and Technology,[3]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

ABSTRACT

Name ambiguity is a special case of identity uncertainty where one person can be referenced by multiple name variations in different situations or even share the same name with other people. In this paper, we present an efficient framework by using two novel topic-based models, extended from Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). Our models explicitly introduce a new variable for persons and learn the distribution of topics with regard to persons and words. Experiments indicate that our approach consistently outperforms other unsupervised methods including spectral and DBSCAN clustering. Scalability is addressed by disambiguating authors in over 750,000 papers from the entire CiteSeer dataset.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Theory

Keywords

Unsupervised Machine Learning, Name Disambiguation.

1. INTRODUCTION

Name queries makes up approximately 5-10% of all searches on the Internet, but they are usually treated by search engines as normal keyword searches without paying attention to the ambiguity of particular names. For example, searching Google for "Yang Song" results in more than 11,000,000 pages, of which even the first page shows five different people's home pages. Beyond the problem of sharing the same name, name misspelling, name abbreviations and other issues compound the challenge of name disambiguation. The same issue also exists in most Digital Libraries (DL), due to the existence of both *synonyms* and *polysems*. In the case of *synonyms*, an author may have multiple name variations in citations, e.g., the author "C. Lee Giles" is sometimes used as "C. L. Giles" in his citations. For *polysems*, different authors may share the same name label in multiple citations, e.g., both "Guangyu Chen" and "Guilin Chen" are used as "G. Chen" in their citations.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. TOPIC-BASED PLSA

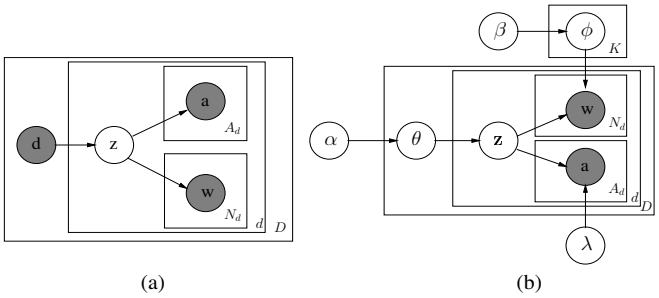

Figure 1: Graphical model representation of (a) PLSA and (b) LDA model. K is the number of topics, D is the total number of documents, N_d is the number of tokens in document d and A_d represents the number of name appearances in document d.

The joint probability of the topic-based PLSA model (see Figure 1(a)) over $d \times a \times w$ is defined as the mixture $P(d,a,w) = P(d)P(a,w|d)$ where $P(a,w|d) = \sum_{z \in Z} P(a,w|z)P(z|d)$.

The definition of the generative model can be described as follows. (1) pick a doc d from the corpus D with probability $P(d)$; (2) select a latent class z_k with probability $P(z_k|d)$; (3) generate a word w with probability $P(w|z_k)$; (4) generate a name a with probability $P(a|z_k)$. Putting it all together, the joint probability can be parameterized by

$$P(d,a,w) = \sum_{z \in Z} P(z)P(z|d)P(w|z)P(a|z). \qquad (1)$$

2.1 Model Fitting with the EM Algorithm

The standard Expectation-Maximization (EM) algorithm is applied to estimate the parameters. In the E-step, we compute $P(z|d,a,w) \propto \frac{P(z)P(a|z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(a|z')P(d|z')P(w|z')}$.

In the M-step, we aim at maximizing the expectation of the complete data likelihood, the formulas are: $P(a|z) \propto \frac{\sum_{d,w} n(d,a,w)P(z|d,a,w)}{\sum_{d,a',w} n(d,a',w)P(z|d,a',w)}$, $P(w|z) \propto \frac{\sum_{a,d} n(d,a,w)P(z|d,a,w)}{\sum_{d,a,w'} n(d,a,w')P(z|d,a,w')}$, $P(z|d) \propto \frac{\sum_{a,w} n(d,a,w)P(z|d,a,w)}{\sum_{d',a,w} n(d',a,w)P(z|d',a,w)}$, where $n(d,a,w)$ denotes the number of occurrences of word w in document d with name a. The EM algorithm stops on convergence, i.e., when the improvement of the log-likelihood is significantly small.

To predict the topics of new documents, $P(w|z)$ are used to estimate $P(a|z)$ for new names a in test document d

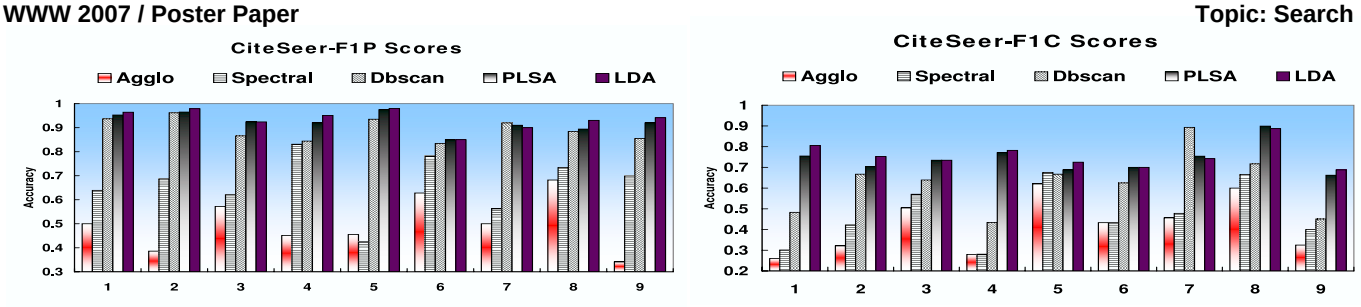

Figure 2: Clustering results on the CiteSeer data set. 1:A. Gupta, 2:A. Kumar, 3:C. Chen, 4:D. Johnson, 5:J. Robinson, 6:J. Smith, 7:K. Tanaka, 8:M. Jones, 9:M. Miller.

through a "folding-in" process [2]. Specifically, the E-step is the same as before; however, the M-step maintains the original $P(w|z)$ and only updates $P(a|z)$ as well as $P(z|d)$.

3. TOPIC-BASED LDA

The generative process of our topic-based LDA model extended from [1] (shown in Figure 1(b)) can be formalized as follows. (1) Draw a multinomial distribution ϕ_z for each topic z from a Dirichlet distribution with prior β; (2) For each document d, draw a multinomial distribution θ_d from a Dirichlet distribution with prior α; (3) For each word w_{di} in d, draw a topic z_{di} from the multinomial distribution θ_d; (4) Draw a word w_{di} from the multinomial distribution $\phi_{z_{di}}$; (5) Draw a name a_{di} from the multinomial distribution $\lambda_{z_{di}}$.

3.1 Inference and Parameter Estimation

The inference problem in LDA is to compute the posterior of the (document-level) hidden variables given a document $d = (\mathbf{w}, \mathbf{a})$ with parameters α and β, i.e., $p(\theta, \phi, \mathbf{z}|\mathbf{w}, \mathbf{a}, \alpha, \beta, \lambda) = \frac{p(\theta,\phi,\mathbf{z},\mathbf{w},\mathbf{a}|\alpha,\beta,\lambda)}{p(\mathbf{w},\mathbf{a}|\alpha,\beta,\lambda)}$. Here $p(\mathbf{w}, \mathbf{a}|\alpha, \beta, \lambda)$ is usually referred to as the marginal distribution of document d: $p(\mathbf{w}, \mathbf{a}|\alpha, \beta, \lambda) = \iint p(\theta|\alpha)p(\phi|\beta)\prod_{n=1}^{N} p(w_n|\theta, \phi)\prod_{m=1}^{M} p(a_m|\theta, \lambda)\, d\theta d\phi$.

By marginalizing over the hidden variable z, the name distribution $p(a|\theta, \lambda)$ can be represented as $\sum_z p(a|z, \lambda)p(z|\theta)$.

As a result, the likelihood of a corpus D can be calculated by taking the product of the marginal probabilities of each of the documents. Specifically, $p(D|\alpha, \beta, \lambda) = \iint \prod_{z=1}^{K} p(\phi_z|\beta)$ $\prod_{d=1}^{N} p(\theta_d|\alpha)\prod_{n=1}^{N} p(w_n|\theta, \phi)\prod_{m=1}^{M} p(a_m|\theta, \lambda)d\theta d\phi$.

To estimate the parameters θ and ϕ, we construct a Markov chain that converges to the posterior distribution on $\mathbf{z}$. The posterior distribution can be derived as follows:

$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{a}) \propto p(z_i = j|\mathbf{z}_{-i})p(w_i|\mathbf{z}, \mathbf{w}_{-i})p(a_i|\mathbf{z}, \mathbf{a}_{-i}) \quad (2)$$

$$\propto \frac{H^{DT}_{dj} + \alpha}{\sum_{j'} H^{DT}_{dj'} + K\alpha} \frac{H^{WT}_{mj} + \beta}{\sum_{m'} H^{WT}_{m'j} + W\beta}, \quad (3)$$

where $\mathbf{z}_{-i}$ means all topic assignments not including the ith word; H^{WT}_{mj} is the number of times word m assigned to topic j except the current instance and H^{DT}_{dj} is the number of times doc d contains topic j except the current instance.

4. EXPERIMENTS

To disambiguate names, we use a hierarchical agglomerative clustering method. Two sets of metrics are applied in our experiments, namely **pair-level pairwise F1** score $F1P$ and **cluster-level pairwise F1** score $F1C$. We also compare with the basic agglomerative clustering, spectral clustering and DBSCAN method.

We collected meta-data from the CiteSeer digital library. Nine most ambiguous author names from the entire data set are tested, the results are shown in Figure 2.

Topic 40 "Database"		Topic 42 "Multimedia"	
query	0.0375	retrieval	0.0411
xml	0.0321	multimedia	0.0411
database	0.0321	broadcast	0.0360
scalability	0.0315	video	0.0311
process	0.0315	shot	0.0311
storage	0.0215	labeling	0.0311
memory	0.0215	flash	0.0215
Jun Yang(Duke)	0.1258	Jun Yang(Duke)	0.0398
Jun Yang(CMU)	0.0477	Jun Yang(CMU)	0.2781

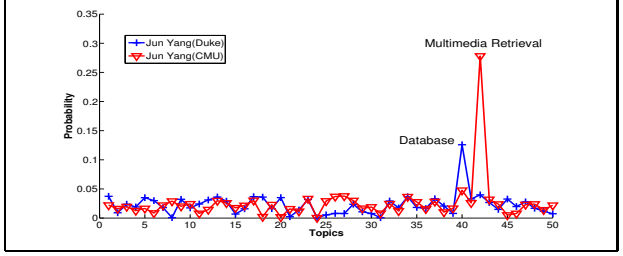

Table 1: LDA results of two different "Jun Yang".

Table 1 lists an illustrative result from LDA. We depict topics that clearly show the differences for disambiguating authors with *exactly* the same name. One "Jun Yang" has very high probability of topic "Database" while the other are highly related with the topic "Multimedia", thus they can be clearly disambiguated from each other.

We empirically tested our models for the entire CiteSeer data set with more than 750,000 documents. PLSA yields 418,500 unique authors in 2,570 minutes, while LDA finishes in 4,390 minutes with 418,775 authors. Considering that our methods only make use of a small portion of the text for each instance (metadata plus the first page), we believe the framework can be efficient for large-scale data sets.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] T. Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, Berkeley, California.