Author Disambiguation: A Nonparametric Topic and Co-authorship Model

Andrew M. Dai
School of Informatics
University of Edinburgh
a.dai@ed.ac.uk

Amos J. Storkey
School of Informatics
University of Edinburgh
a.storkey@ed.ac.uk

Abstract

A fully generative model is provided for the problem of author disambiguation. This approach infers the topics for each author and combines that with co-author information. The problems involved are similar to other entity resolution problems where differing references may refer to one author entity and identical references may refer to different author entities. We extend the hierarchical Dirichlet process and nonparametric latent Dirichlet allocation models to tackle this problem in a nonparametric, generative manner making no prior assumptions on the number of author entities, topics or research groups in the corpus. The model develops a hierarchical Dirichlet process for author-topic combinations. It conditions this model at document level on another hierarchical Dirichlet process for research groups. This enables the authors and topics to be suitably coupled. We perform joint inference to sample the author entities, topics and their group memberships. We present results from our approach on real-world datasets.

1 Introduction

Entity resolution or record linkage is a difficult problem that is often one of the first steps in data mining to reduce the noise in a dataset. When applied to authors and citation databases, this becomes a problem of discovering real-world author entities in the absence of unique identifiers. Email addresses and author institutions can be used to help with the likely identity of an author name. However these are often vulnerable to change when the author's circumstances change. Author names are also liable to have multiple variants, either through transliteration, transcription errors, OCR errors or spelling mistakes. These errors can also be passed on in citations.

Traditional ER solutions often require formulating rules, are rarely generative models and have results that can be hard to interpret probabilistically. Models which use information from other fields also perform better than those that solely use names [1]. Other models have also integrated topic information, e.g. [2], which identifies the topics that authors frequently write on. However, their method is tailored for knowledge discovery rather than author disambiguation and is unable to take advantage of co-author information.

A nonparametric Bayesian approach to the problem is followed, based on the hierarchical Dirichlet process [3]. The model developed in this work involves two hierarchical Dirichlet processes, one conditioned on the other at document level. The first of these is a joint model over author-entities and topics, which is used to model both words and author records by embedding them in a joint author-topic space. The second is a non-parametric latent Dirichlet allocation model, used to associate data-items with particular research groups. In this way, authors (and the topics they are associated with) become linked with co-authors.

This work develops a number of novel methods for entity resolution, and novel approaches for combining hierarchical Dirichlet processes. One important impact of this work is the combination of the nonparametric latent Dirichlet allocation and product space hierarchical Dirichlet topic models. However the most important achievement is the development and application of the full topic and co-authorship model for citation data, with its ability to deal with multiple entity types within an individual document, and fully automated group size and topic size inference.

1.1 Hierarchical Dirichlet Processes

The hierarchical Dirichlet process is a hierarchical extension to the Dirichlet process [3] and models a hierarchical dependency between multiple Dirichlet processes that aims to share clusters among groups of data. It uses a base measure for a set of Dirichlet processes that is itself distributed according to a Dirichlet process. This defines a set of probability measures G_i for N groups and a global probability measure G_0, which is itself distributed as a Dirichlet process with a base measure H and concentration parameter γ. Each G_i is used as a base measure for another Dirichlet Process with concentration parameter τ. We write $G_0|\gamma, H \sim \mathrm{DP}(\gamma, H)$, and $G_i|\alpha_0, G_0 \sim \mathrm{DP}(\tau, G_0)$.

2 The Author, Topic and Research Group Model

This work models the words and author records in documents via the use of latent author and topic entities, and latent research groups. The topics and author names are modelled in a joint space. Topic entities are distributions over the vocabulary of words in the corpus and author entities are distributions over representations of an author's real name. From this, an author-topic entity is composed of both of these types of entities. Each author-topic entity is a member of one or more research groups. The hierarchical Dirichlet process over research groups will imply that if one author in a particular research group is named on a paper, then other authors within the same group are more likely to be co-authors.

To integrate co-authors into the model, we use a hierarchical stick-breaking process to represent the research groups in the corpus. The distribution of authors and words is then conditional on the research group that they are assigned to, so that we arrive at the joint distribution for a document as $P(\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{g}) = P(\boldsymbol{g})P(\boldsymbol{z}|\boldsymbol{g})P(\boldsymbol{a}|\boldsymbol{z})P(\boldsymbol{w}|\boldsymbol{z})$ where $\boldsymbol{w}$ denote the words for a document, $\boldsymbol{a}$ denote the authors for a document, $\boldsymbol{z}$ denote the author-topic entity allocations and $\boldsymbol{g}$ denote the research group allocations. The full generative model is presented in Figure 1.

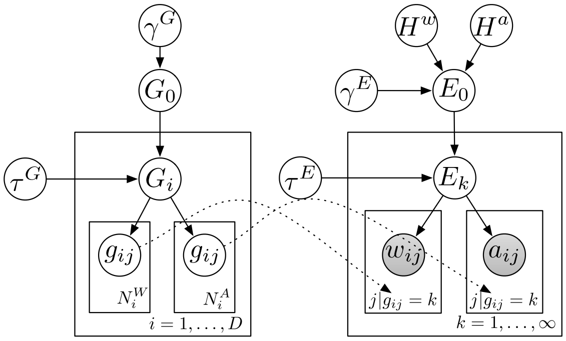

Figure 1: Our generative model in plate notation. The left side denotes the group structure and the right side denotes the entity structure given the group assignments g. i ranges over documents, j ranges over the author names and the words in the document and k ranges over the research groups. H denotes the base measures. G denotes the Dirichlet processes for the groups. E denotes the Dirichlet process for the author-topic entities.

2.1 Research Group Model

At corpus level, the probabilities associated with each research group is sampled from a stick breaking construction to form a base distribution. Then for each document, each data item (word or author) in the document is associated with each research group using a Dirchlet process.

Formally, the corpus-level Dirichlet process $G_0 \sim \mathrm{GEM}(\gamma^G)$ is used to represent the prior over the entirety of research groups. There is also a Dirichlet process for each document $G_i \sim \mathrm{DP}(\tau^G, G_0)$ that represents a prior over the research groups that appear in a document. The use of a hierarchical group structure results in a posterior where the data items in a document are likely to appear in the same research group. At the datapoint level, each datapoint (word or author) samples a group $g_{ij} \sim G_i$. The author entities are sampled for each research group conditional on this group structure.

2.2 Entity Model

At the top of the model, there is a corpus-level Dirichlet process over author-topic entities. Each entity defines the topic and a canonical author name. The distribution over author topic entities provide a base distribution for each research group. The corpus level Dirichlet process is given by $E_0 \sim \mathrm{DP}(\gamma^E, H^w \times H^a)$. This base distribution is the product of the Dirichlet distribution H^w for the topic parameters and uniform distribution H^a for the canonical authors. Each data item in the research group is then associated with an author-topic entity using a Dirichlet process with the aforementioned base distribution. The Dirichlet process in each research group is denoted $E_i \sim \mathrm{DP}(\tau^E, E_0)$, and is a prior over the author entities that appear in that research group.

Finally, if the data item is an author name, the author name variant is then sampled from the canonical name according to a generative name corruption model [4]. If the data item is a word, the word is sampled from that word distribution associated with the given topic.

1. Sample the global group distribution $\boldsymbol{\pi}_0^G \sim \mathrm{GEM}(\gamma^G)$
2. Sample the global entity distribution $\boldsymbol{\pi}_0^E \sim \mathrm{DP}(\gamma^E, H^w \times H^a)$
3. For each group k, sample the group entity distribution $\boldsymbol{\pi}_k^E \sim \mathrm{DP}(\tau^E, \boldsymbol{\pi}_0^E)$
   (a) For each member of k, sample an author-topic entity $z_k^G \sim \boldsymbol{\pi}_k^E$
4. For each author-topic entity e:
   (a) Sample the topic parameters $\phi_e^w \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   (b) Sample the author canonical name $\phi_e^a \sim \mathrm{Unif}(1, \ldots, a_N)$
5. For each document i, sample $\boldsymbol{\pi}_i^G \sim \mathrm{DP}(\tau^G, G_0)$
   (a) For each data item $j = 1, \ldots, N_i$
      i. Sample the group $g_{ij} \sim \boldsymbol{\pi}_i^G$
      ii. Sample the author-topic entity $z_{ij} | g_{ij} \sim \boldsymbol{\pi}_{g_{ij}}^E$
      iii. Sample the word $w_{ij} | z_{ij} \sim \mathrm{Mult}(\phi_{z_{ij}}^w)$ or the author $a_{ij} | z_{ij} \sim f^a(\phi_{z_{ij}}^a)$

Inference is performed with collapsed Gibbs sampling [3]. We use a modified version of Algorithm 8 of Neal's algorithms [5] to sample the non-conjugate parameters for the name model since having a discrete base measure results in multiple classes likely having the same parameters so slowing mixing. We sample the canonical names from one of the names currently assigned to the entity after sampling the research group and entity allocations.

3 Experiments

We evaluate our approach on the hand-labelled real-world citation databases: CiteSeer, arXiv (HEP) and Rexa. The CiteSeer dataset was originally created by Giles et al. [6]. The HEP dataset was used in the KDD Cup 2003 competition. The Rexa dataset was obtained from Aron Culotta[1]. We use CiteSeer and HEP datasets that have been cleaned by Bhattacharya et al.[2]. The CiteSeer, HEP and Rexa datasets contain 2,892, 58,515 and 1,972 references respectively. For all datasets, we applied a stoplist to the abstracts and removed words which appeared less than 10 times. This resulted in a vocabulary size for the Citeseer, HEP and Rexa datasets of 483, 6,155 and 694 respectively. Where the abstracts could not be found in the CiteSeer and Rexa datasets, we used the paper titles instead.

We put vague Gamma priors on the concentration parameters and sample the parameters. These methods were also tested using other name corruption models (an overlapping multinomial trigram model and a nonparametric extension of the generative bigram model [7]). Though the performance of these models was reduced, they still showed significant capability of utilising coauthor information to disambiguate authors. We implemented the Author-Topic model [2] and augmented their model with our non-conjugate generative name model. We tried different values of T, the number of author-topic entities, set to values around the real number of authors in the datasets. T_1, T_2, T_3 are set to 1000, 800, 600 respectively for CiteSeer and 800, 600, 400 for Rexa. Finally, we ran a very basic baseline model which merged identical names together. The sampling converges after around 100 iterations taking around 4s per iteration for the Citeseer dataset. It can be seen from Table 1 that the modelling of research groups and coauthors contributes much more to performance

[1] http://www.cs.umass.edu/∼culotta/data/rexa.html

[2] http://www.cs.umd.edu/projects/linqs/projects/er/index.html

Table 1: Pairwise disambiguation recall, precision and F1 results for runs with and without research groups and topics and the LDA author topic model. The last 10 samples of each of 10 parallel runs were averaged together. A burn-in of 500 was used.

	Rexa			**Citeseer dataset**			**HEP**		
Model	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Groups + Topics	0.891	0.999	0.942	0.966	0.985	0.976	0.976	0.944	0.960
Groups - Topics	0.963	0.997	0.980	0.961	0.994	0.977	0.966	0.986	0.976
+ Topics	0.971	0.997	0.984	0.984	0.895	0.937	0.935	0.982	0.958
- Topics	0.970	0.998	0.984	0.982	0.929	0.955	0.969	0.966	0.967
Author-Topic (T_1)	0.447	0.995	0.617	0.678	0.918	0.780	-	-	-
Author-Topic (T_2)	0.776	0.995	0.872	0.814	0.800	0.835	-	-	-
Author-Topic (T_3)	0.810	0.994	0.892	0.882	0.692	0.775	-	-	-
Base	0.526	0.999	0.689	0.579	0.999	0.689	-	-	-

Table 2: Examples of inferred research groups, with associated names and topics

Names	**Topics**	**Names**	**Topics**
Robyn Kozierok, R. Kozierok	agent, hypermedia	T. Jaakkola	environments, stochastic,
Pattie Maes, P. Maes	adaptive, www	Singh, S.	policies, discrete, markov

than solely modelling topics, however the addition of topic information in these cases seem to reduce performance. Since the topics in our datasets are relatively homogenous and from small domains we would expect better differentiation of authors and topics with a broader dataset or more ambiguous author references. Since we also couldn't locate the complete abstracts for some of the documents this may also have impacted performance. An example of an entity that was disambiguated correctly is *Reisbeck, C. K.* and *Reisbech, C.* Our model performs much better than the author-topic model [2] for the problem of author disambiguation though again, this may be due to the topics in the document overwhelming the authors. Our models performance is slightly less than in [4] primarily through the difference in a more broadly specified name noise model. Their method also requires the number of groups to be specified in advance and this effects their Recall/Precision tradeoff. Finally their method assumes that authors with identical names refer to the same real entity. Their datasets have few ambiguous authors so this doesn't significantly affect precision.

4 Conclusion

We present a model to disambiguate authors in a corpus of documents. The model incorporates information from the co-authors for each paper to model research groups and models the topics and their corresponding authors jointly for each paper. The model shows significant improvement over ignoring research groups. The model is fully automated in that it does not require pre-specification of numbers of research groups, topics etc. Because it can use either topic information, group information or both to aid the resolution it is versatile in the situations it can be employed in.

References

[1] Rob Hall, Charles Sutton, and Andrew Mccallum. Unsupervised deduplication using cross-field dependencies. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 310–317, New York, NY, USA, 2008. ACM.

[2] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[4] Indrajit Bhattacharya and Lise Getoor. A latent Dirichlet model for unsupervised entity resolution. In *The SIAM International Conference on Data Mining (SIAM-SDM)*, Bethesda, MD, USA, 2006.

[5] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[6] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, 1998.

[7] Hanna M. Wallach. Topic modeling: beyond bag-of-words. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 977–984, New York, NY, USA, 2006. ACM.