A Labeled LDA Approach to Understanding the Dynamics of Collaboration

Nikhil Johri
Department of Computer Science
Stanford University
Stanford, CA, USA.
njohri2@stanford.edu

Abstract

We propose a topic modeling approach to understand the nature of academic collaborations between individuals. Specifically, we use Labeled LDA (Ramage et al., 2009), a variation of the popular topic model Latent Dirichlet Allocation (Blei et al., 2003), to train a set of *author-specific* topics over the ACL corpus. The ACL corpus ranges from 1965 to 2009, and we train a separate topic model for each year over the papers published upto the given year. Once we have trained these models, we examine the influence present in a publication from each of its authors. We use a function of the cosine similarity score between the document's term vector and each author's topic signature in the year preceding the document's publication as our metric. We suggest a theory of the different types of academic collaboration prevalent and discuss how our system performs at classifying high impact papers into these types. Finally, we demonstrate how our system can be applied to answer several questions regarding the nature of collaborations in Computational Linguistics research.

1 Introduction

The understanding of academic collaboration has attracted much interest in the realm of the social sciences, and is recently gaining traction in computer science, particularly from the viewpoint of social network analysis.

We propose a theoretical framework for determining the types of collaboration present in a document, based on factors such as the number of established authors, the presence of unestablished authors and the similarity of the established authors' past work to the document's term vector. These collaboration types attempt to describe the nature of co-authorships between of students and advisors as well as those solely between established authors in the field. We present a decision diagram for classifying papers into these types, as well as a description of the intuition behind each collaboration class.

Once we have a proposed theory, we attempt to create a system that can automatically categorize collaborative works into their collaboration types. Labeled Latent Dirichlet Allocation (LLDA) (Ramage et al., 2009) is used to realize this task. LLDA has found success in a number of natural language processing topic modeling tasks, such as the *credit attribution* problem. For our system, we use LLDA to train topic models over the ACL corpus for every year. Each topic, in this case, pertains to a single author. Using the author signatures so obtained, we come up with a series of metrics to determine how we might classify each document.

We qualitatively analyze our results by examining the categorization of several high impact papers. With consultation from prominent researchers and textbook writers in the field, we verify how accurate these results are. However, given the subjective nature of the categorization and collaboration types, it is hard to come up with a reasonable quantitative evaluation.

2 Related Work

In recent years, popular topic models such as LDA (Blei et al., 2003) have been increasingly used to study the history of science by observing the changing trends in term based topics (Hall et al., 2008), (Gerrish and Blei, 2010). In the case of Hall et al., regular LDA topic models were trained over the ACL anthology on a per year basis, and the changing trends in topics were studied from year to year. Gerrish and Blei's work computed a measure of influence by using Dynamic Topic

Models (Blei and Lafferty, 2006) and studying the change of statistics of the language used in a corpus.

These models propose interesting ideas for utilizing topic modeling to understand of scientific history. However, our primary interest, in this paper, is the study of academic collaboration between different authors; we therefore look to learn models for authors instead of only documents. Popular topic models for authors include the Author-Topic Model (Rosen-Zvi et al., 2004), a simple extension of regular LDA that adds an additional author variable over the topics. The Author-Topic Model learns a distribution over words for each topic, as in regular LDA, as well as a distribution over topics for each author. Alternatively, Labeled LDA (Ramage et al., 2009), another LDA variation, offers us the ability to directly model authors as topics by considering them to be the topic labels for the documents they author.

In this work, we use Labeled LDA to directly model probabilistic term 'signatures' for authors. As in (Hall et al., 2008) and (Gerrish and Blei, 2010), we learn a new topic model for each year in the corpus, allowing us to account for changing author interests over time.

3 Methodology

3.1 Labeled Latent Dirichlet Allocation

Latent Dirichlet Allocation, or LDA (Blei et al., 2003), is a widely popular technique of probabilistic topic modeling where each document in a corpus is modeled as a mixture of 'topics', which themselves are probability distributions over the words in the vocabulary of the corpus. LDA is completely unsupervised, assumes that a latent topic layer exists and that each word is generated from one underlying topic from this set of latent topics.

Labeled Latent Dirichlet Allocation, or LLDA (Ramage et al., 2009), is a variation on the regular LDA topic model whereby a one-by-one relation is defined between topics and tags (or in our domain, authors). This constrains each topic to correspond to exactly one author when learning the model. As a result, we retrieve for each author a direct distribution over terms. This distribution serves as a 'signature' for an author, dominated by the terms frequently used by the author. It is useful to note that the advantage of using Labeled LDA over another topic model comes in the fact that topics are constrained to pre-assigned labels. This assures us that each topic created will correspond to a single author, unlike regular LDA where terms cluster based on semantic topics.

We train a separate LLDA model for each year in the corpus, training on only those papers written before and during the given year. Thus, we have separate 'signatures' for each author for each year, and each signature only contains information for the specific author's work upto and including the given year.

3.2 Types of Collaboration

There are several ways one can envision to differentiate between types of academic collaborations. We focus on three factors when creating collaboration labels, namely:

- Presence of unestablished authors
- Similarity to established authors
- Number of established authors

If an author whom we know little about is present on a collaborative paper, we consider him or her to be a new author. We threshold new authors by the number of papers they have written upto the publication year of the paper we are observing. Depending on whether this number is below or above a threshold value, we consider an author to be *established* or *unestablished* in the given year.

Similarity scores are measured using the trained LLDA models described in Section 3.1. For any given paper, we measure the similarity of the paper to one of its (established) authors by calculating the cosine similarity of the author's signature in the year preceding the paper's publication to the paper's term-vector.

Using the aforementioned three factors, we define the following types of collaborations:

- **Apprenticeship Paper** This describes a paper which is authored by one or more established authors and one or more unestablished authors, such that the similarity of the paper to more than half of the established authors is high. In this case, we say that the new author (or authors) was an apprentice of the established authors, continuing in their line of work.

- **New Blood Paper** This describes a paper which is authored by one established author and one or more unestablished authors, such that the similarity of the paper to the established author is low. In this case, we say that the new author (or authors) provided new ideas or worked in an area that was dissimilar to that which the established author was working in.

- **Synergistic Paper** This describes a paper authored only by established authors such that it does not heavily resemble any authors' previous work. In this case, we consider the paper to be a product of synergy of its authors.

- **Catalyst Paper** This is similar to a Synergistic Paper, with the exception that unestablished authors are also present on a Catalyst Paper. In this case, we hypothesize that the unestablished authors were the catalyst responsible for getting the established authors to work on a topic unlike their previous work.

- **Pollinator Paper** This is similar to a Catalyst Paper, with the exception that the paper does resemble one of the established authors. In this case, we say that the unestablished authors acted as 'pollinators' in bringing the work of the solitary established author with a high similarity score to the other established authors.

The decision diagram in Figure 1 presents an easy way to determine the collaboration type assigned to a paper.

4 Preliminary Results

Following the decision diagram presented in Figure 1 and using similarity scores based on the values returned by the LLDA models (Section 3.1, we can deduce the collaboration type to assign to any given paper. However, absolute categorization requires an additional thresholding of author similarity scores. To avoid the addition of an arbitrary threshold, instead of directly categorizing papers, we rank them based on the calculated similarity scores. We consider the following three scores, and present examples of how high impact papers (based on page rank, as calculated by (Radev et al., 2009)) fall into the various spectra:

4.1 The Apprentice-New Blood Score

This score corresponds to papers with a single established author and one or more unestablished authors (presumably encapsulating various advisor-student collaborations). Numerically, we measure this score as the cosine similarity between the established author's signature and the paper, $cos(a_{sig}, paper)$. The higher this score, the more the paper is like an Apprenticeship Paper (as it resembles the established author or advisor), while a score closer to the lower end of the range is more characteristic of a New Blood Paper.

4.1.1 Example: Apprenticeship Paper

Improvements in Phrase-Based Statistical Machine Translation (2004)
by Richard Zens and Hermann Ney
This paper had a high Apprentice-New Blood score, indicating high similarity to established author Hermann Ney. This categorizes the paper as an Apprenticeship Paper.

4.1.2 Example: New Blood Paper

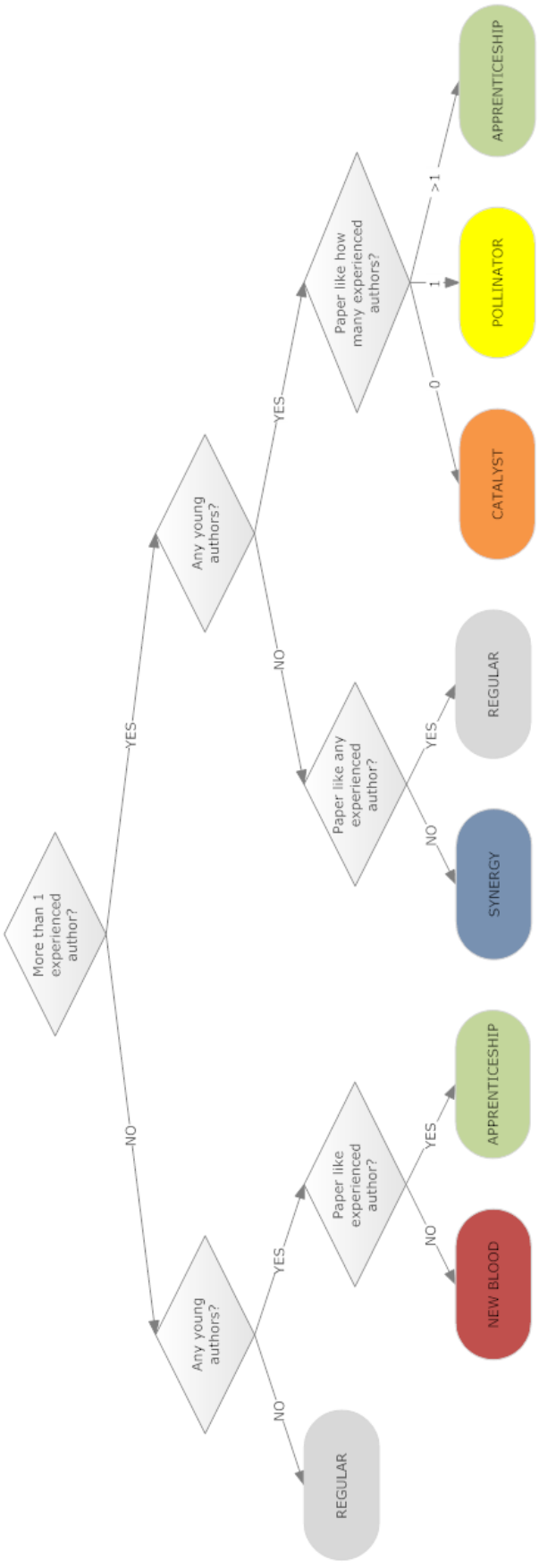
Thumbs up? Sentiment Classification using Machine Learning Techniques (2002)
by Lillian Lee, Bo Pang and Shivakumar Vaithyanathan
This paper had a low Apprentice-New Blood score, indicating low similarity to established author Lillian Lee. This categorizes the paper as a New Blood Paper, with new authors Bo Pang and Shivakumar Vaithyanathan. It is important to note here that new authors do not necessarily mean young authors or grad students; in this case, the final author was established, but in a field outside of ACL.

4.2 The MaxSim Score

This score describes the maximum similarity, or the minimum deviation present in a paper with multiple established authors, i.e. $max_{a \in est} cos(a_{sig}, paper)$. On papers with only established authors, this helps differentiate high synergy papers from low synergy papers - papers with lower MaxSim scores have higher synergy and vice versa. If unestablished authors are present, this helps determine whether or not the paper is a Catalyst Paper.

4.2.1 Example: High Synergy Paper

Catching the Drift: Probabilistic Content Models, with Applications to Generation and Sum-



.

Figure 1: Decision diagram for determining the collaboration type of a paper. A minimum of 1 established author is assumed.

marization (2003)
by Regina Barzilay and Lillian Lee
This paper had low similarity to both established authors on it, making it a highly synergistic paper. Synergy here indicates that the work done on this paper was mostly unlike work previously done by either of the authors.

4.2.2 Example: Catalyst Paper

Answer Extraction (2000)
by Steven Abney, Michael Collins, Amit Singhal
This paper had a very low MaxSim score, as well as the presence of an unestablished author, making it a Catalyst Paper. The established authors (from an ACL perspective) were Abney and Collins, while Singhal was from outside the area and did not have many ACL publications. The work done in this paper focused on information extraction, and was unlike that previously done by either of the ACL established authors. Thus, we say that in this case, Singhal played the role of the catalyst, getting the other two authors to work on an area that was outside of their usual range.

4.3 The Apprentice-Pollinator Score

This score helps differentiate between Pollinator and Apprenticeship papers when multiple established authors and one or more unestablished author are present on a paper. If the MaxSim score for such a paper is high, it means that there is an established author whose previous work resembles this paper. If this is the case, we find the difference between that author's similarity score and the next highest similarity score for an established author on the paper. If this value is high, we consider the unestablished authors to be playing the role of pollinators for the high similarity score author; if it is low, we consider them to be apprentices of the established authors.

5 Applications

A number of questions about the nature of collaborations may be answered using our system. We describe approaches to some of these in this section.

5.1 The Hedgehog-Fox Problem

From the days of the ancient Greek poet Archilochus, the Hedgehog-Fox analogy has been frequently used (Berlin, 1953) to describe two different types of people. Archilochus stated that "The fox knows many things; the hedgehog one big thing." A person is thus considered a 'hedgehog' if he has expertise in one specific area and focuses all his time and resources on it. On the other hand, a 'fox' is a one who has knowledge of several different fields, and dabbles in all of them instead of focusing heavily on one.

Author	Avg. Sim. Score
Koehn, Philipp	0.43456
Pedersen, Ted	0.41146
Och, Franz Josef	0.39671
Ney, Hermann	0.37304
Sumita, Eiichiro	0.36706
Zhang, Min	0.36516
Vogel, Stephan	0.36498
Satta, Giorgio	0.35890
Strzalkowski, Tomek	0.35862
Moore, Robert C.	0.34960

Table 1: Hedgehogs - authors with the highest average similarity scores

We show how, using our system, one can discover the hedgehogs and foxes of Computational Linguistics. We look at the top 100 published authors in our corpus, and for each author, we compute the average similarity score the author's signature has to each of his or her papers. Note that we start taking similarity scores into account only after an author has published 5 papers, thereby allowing the author to stablize a signature in the corpus and preventing the signature from being boosted by early papers (where author similarity would be artificially high, since the author was new).

We present the authors with the highest average similarity scores in Table 1. These authors can be considered the hedgehogs, as they have highly stable signatures that their new papers resemble. On the other hand, Table 2 shows the list of foxes, who have less stable signatures, presumably because they move about in different areas.

5.2 Similarity to previous work by sub-fields

Based on the different types of collaborations discussed in, a potential question one might ask is which sub-fields are more likely to produce *apprentice* papers, and which will produce *new blood* papers. To answer this question, we first need to determine which papers correspond to which subfields. Once again, we use topic models to solve this problem. We first filter out a subset of the

Author	Avg. Sim. Score
Marcus, Mitchell P.	0.09996
Pustejovsky, James D.	0.10473
Pereira, Fernando C. N.	0.14338
Allen, James F.	0.14461
Hahn, Udo	0.15009
Wilks, Yorick	0.15526
Weischedel, Ralph M.	0.16211
Su, Jian	0.17071
Hovy, Eduard H.	0.17402
Lin, Dekang	0.17620

Table 2: Foxes - authors with the lowest average similarity scores

1,200 highest page-rank collaborative papers from the years 1980 to 2007. We use a set of topics built by running regular LDA over the ACL corpus, in which each topic is hand labeled by experts based on the top terms associated with it. Given these topic-term distributions, we can once again use the cosine similarity metric to discover the highly associated topics for each given paper from our smaller subset, by choosing topics with cosine similarity above a certain threshold δ (in this case 0.1).

Once we have created a paper set for each topic, we can measure the 'novelty' score for each paper by finding:

$max_{a \in est} cos(a_{sig}, paper)$

where est is the set of experienced authors, i.e. authors who have written $\geq X$ paper for some threshold X, which varies depending on the year. We use the author's signature, a_{sig}, from the year preceeding the publication of the paper when measuring author-paper similarity. We choose to observe the similarity scores only for established authors as newer authors will not have enough previous work to produce a stable term signature, and we vary the threshold by year to account for the fact that there has been a large increase in the absolute number of papers published in recent years.

Once we have computed these values, we can find the average score for each topic. This average similarity score gives us a notion of how similar to the established author (or authors) a paper in the sub field usually is. Low scores imply indicate that new blood and synergy style papers are more common, while higher scores imply more non-synergistic or apprenticeship style papers. This could indicate that topics with lower scores are more open ended, while those with higher scores require more formality or training. The top five topics in each category are shown in Tables 3 and 4. The scores of the papers from the two tables were compared using a t-test, and the difference in the scores of the two tables was found to be very statistically significant with a two-tailed p value $<< 0.01$.

Topic	Score
Prosody	0.2341
Unification Based Grammars	0.2236
Bilingual Word Alignment	0.2197
Categorial Grammar + Logic	0.2195
Statistical Machine Translation	0.2190

Table 3: Topics with highest established author similarity scores

Topic	Score
Question Answering Dialog System	0.1283
Sentiment Analysis	0.1465
Summarization	0.1496
Planning/BDI	0.1505
Anaphora Resolution	0.1558

Table 4: Topics with lowest established author similarity scores

6 Conclusion

In this paper, we have described a theory whereby multi-author academic papers in Computational Linguistics can be categorized into different collaboration types based on certain characteristics. We used topic modeling, a very popular technique in the NLP community, to realize this theory, by creating yearly 'author term signatures' which were then used to measure an author's similarity to a paper.

There were several design choices we could use to implement our system. We chose to use Labeled LDA as it allowed us the ability to directly model a probabilistic term signature for individual authors. An alternative topic model like the Author-Topic Model would have been appropriate as well, however, it would have added an additional topic layer between authors and terms, which Labeled LDA avoids. Simpler approaches like regular TF-IDF of terms could also be considered, but topic modeling adds the sophistication of a generative, probabilistic approach wherein a simpler technique would fall short.

Once we had a system that categorized papers as different collaboration types, we claimed that it could be used to answer several questions about the nature of academic collaborations in Computational Linguistics. We demonstrated it's utility in answering two such questions via simple alterations. Using similar modifications, there are a number of further questions that this system will be able to answer, such as:

- Are authors that start out as new blood authors more successful later on than those that start as apprentices?
- Do certain types of collaboration benefit certain subfields more than others?
- Is there any influence of gender on collaboration type?

In the future, it would be useful to have a quantitative evaluation for the model. We are presently in the process of getting senior researchers and textbook authors in the field of Computational Linguistic and Language Processing to provide verification for our results by indicating whether or not a collaboration categorization is accurate for a paper. This verification will be useful in understanding how well our model works, and would also allow us to suitable adjust our threshold values.

Acknowledgments

I would like to thank Professor Dan Jurafsky, Professor Dan McFarland and Dan Ramage for their valuable feedback and guidance throughout this project.

References

Isaiah Berlin. 1953. *The hedgehog and the fox: An essay on Tolstoy's view of history*. Simon & Schuster.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Sean M. Gerrish and David M. Blei. 2010. A language-based approach to measuring scholarly impact. In *Proceedings of the 26th International Conference on Machine Learning*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The acl anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 54–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494.