Finding Identity Group "Fingerprints" in Documents

Lashon B. Booker

The MITRE Corporation, 7515 Colshire Drive,
McLean, Virginia 22102 USA
booker@mitre.org

Abstract. This paper describes how social identity group "fingerprints" can be extracted from a document collection by applying topic analysis methods in a novel way. The results of document classification experiments suggest that these group-level attributes provide better predictions of group affiliation than document-level attributes. Applications of this method for forensic authorship analysis are also discussed.

Keywords: topic analysis, latent dirichlet allocation, document classification, authorship analysis.

1 Introduction

The identity of an individual or target group responsible for authoring a text document or message can be a critical piece of evidence in many criminal investigations [10]. Computational approaches to authorship analysis usually focus on structural characteristics and linguistics patterns in a body of text [3]. While these approaches provide some important forensic capabilities, there remains a need for some way to discern the ideas and intentions conveyed in the text and use those qualities to help determine authorship [10].

Recently developed text analysis techniques may offer a feasible way to automatically compute such a semantic representation of text. Generative probabilistic models of text corpora, such as Latent Dirichlet Allocation, use mixtures of probabilistic "topics" to represent the semantic structure underlying a document [1]. Each topic is a probability distribution over words and the gist or theme of a document is represented as a probability distribution over those topics. Studies suggest that topic models give a better account of the properties of human semantic memory than latent semantic analysis models which represent each word as a single point in a semantic space [5].

When considering how to use this capability for authorship analysis, it is important to recognize that many factors influence the ideas present in a document, even when that document has just a single author. In particular, factors related to social identity - such as age, gender, ideology, beliefs, etc. – play an important role in communication behaviors. If the attributes of these factors could be teased out of a document, they might provide a valuable "fingerprint" facilitating author analysis. This paper describes preliminary experiments on a computational approach to extracting such identity group fingerprints.

Our investigation of these ideas begins by focusing on the role of identity groups in the formation of collaborative networks associated with scientific publications. This is a good starting point because many of the social factors influencing the content of a scientific publication can be readily identified and there are large amounts of publically available data to work with. We describe how to characterize identity groups and compute identity group attributes in this domain. We also present experiments showing that the group attributes provide better predictions of the contents of documents published by group members than attributes derived from the individual documents. Finally, we demonstrate how the techniques developed to extract identity group attributes can be used more broadly for document classification in general.

2 Identity Groups and Scientific Collaboration

An important goal of studies examining the collaborative networks associated with scientific publications is to understand how collaborative teams of co-authors are assembled. What are the important social identity groups that influence the way teams are assembled to write a paper in this setting? Publication venues are visible manifestations of some of those key social identity groups. People tend to publish papers in conferences where there is some strong relationship between their interests and the topics of the conference. Consequently, conference participants tend to have overlapping interests that give them a meaningful sense of social identity.

There are many attributes that might be useful for characterizing these social identity groups, ranging from the various social and academic relationships between individuals to the organizational attributes of professional societies and funding agencies. One readily available source of information about a group is the corpus of documents that have been collectively published by group members. Given such a document collection as a starting point, topics and frequently used keywords are an obvious choice as identity group attributes in this domain. The peer review system is a mechanism that ensures published papers include enough of the topics and keywords considered acceptable to the group as a whole. Authors that do not conform to these norms and standards have difficulty getting their papers published, and have difficulty finding funding for their work.

2.1 Topic Analysis

If topics are considered to be the identity group attributes of interest, it is natural to turn to topic analysis as a way of identifying the attributes characterizing each group. Topic analysis techniques have proven to be an effective way to extract semantic content from a corpus of text. Generative probabilistic models of text corpora, such as Latent Dirichlet Allocation (LDA), use mixtures of probabilistic “topics” to represent the semantic structure underlying a document [1]. Each topic is a probability distribution over the words in the corpus. The gist or theme of a topic is reflected in selected words having a relatively high probability of occurrence when that topic is prevalent in a document. Each document is represented as a probability distribution over the topics associated with the corpus. The more prevalent a topic is in a document, the higher its relative probability in that document's representation. We use an LDA model of the scientific document collections in our research.

An unsupervised Bayesian inference algorithm can be used to estimate the parameters of an LDA model; that is, to extract a set of topics from a document collection and estimate the topic distributions for the documents and the topic-word distributions defining each topic. As illustrated in Figure 1, the algorithm infers a set of latent variables (the topics) that factor the word document co-occurrence matrix. Probabilistic assignments of topics to word tokens are estimated by iterative sampling. See [4] for more details. Once the model parameters have been estimated, the topic distribution for a new document can be computed by running the inference algorithm with the topic-word distribution fixed (i.e., use the topic definitions that have already been learned).

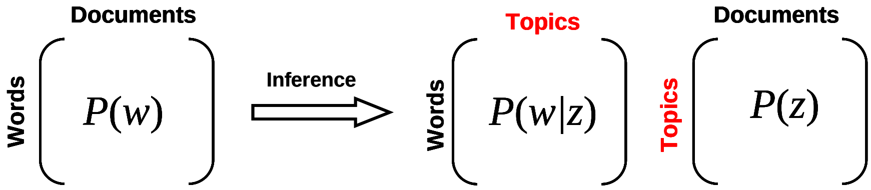


Fig. 1. Inferring the parameters of an LDA topic model.

2.2 Document Classification

If topics are indeed identity group attributes for publication venues viewed as social identity groups, then these attributes ought to distinguish one group from another somehow. Different groups should have distinguishable topic profiles and the topic profile for a document should predict its group. Document classification experiments can be used to verify that identity group influence on published papers is reflected in document topic profiles.

The document descriptions derived from an LDA model are ideally suited to serve as example instances in a document classification problem. One outcome of estimating the parameters of an LDA model is that documents are represented as a probability distribution over topics. Each topic distribution can be interpreted as a set of normalized real-valued feature weights with the topics as the features. Results in the literature suggest that these features induced by an LDA model are as effective for document classification as using individual words as features, but with a big advantage in dimensionality reduction [1].

Regardless of what features we use, we would expect that documents from different topic areas would be distinguishable in a document classification experiment. A more interesting question is whether or not one set of features provides more discriminatory information than another. This is where our research hypothesis about group-level attributes becomes relevant. We hypothesize that in many cases the group identity is most effectively represented by group-level attributes. What are group-level attributes associated with the document collections being considered here?

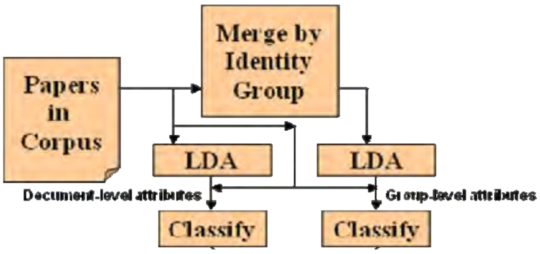


Fig.2. Two ways to generate topic-based attributes for classifying documents

The standard approach to using LDA is to compute topics for the entire corpus that account for the word co-occurrences found in each individual document. We view these topics as document-level attributes. Our modeling emphasis on identity groups led us to consider an alternative approach that focuses instead on word co-occurrences at the group level. A very simple procedure appears to make this possible, as illustrated in Figure 2. First we aggregate the documents affiliated with an identity group into a single mega-document. Then, we use LDA to compute topics for the resulting collection of mega-documents. Since the topics computed in this way account for word co-occurrences in the mega-documents, they are attributes of the group and not attributes associated with the individual documents. However, these group-level topics can also be used as attributes to characterize each individual document, since all topics are represented as probability distributions over the same vocabulary of words. We hypothesize that the topics computed in this manner directly capture the attributes associated with each identity group as a whole and in some sense should be better than attributes derived from the word co-occurrences in individual documents.

2.3 Document Classification Experiments

In order to test this hypothesis, we conducted a document classification experiment in which topics extracted from an unlabeled body of text were used to predict the social identity group (that is, publication venue) of that document. The document collection for this experiment was selected from a dataset containing 614 papers published between 1974 and 2004 by 1036 unique authors in the field of Information Visualization [2]. The dataset did not include the full text for any of the documents, so we worked with the 429 documents in the dataset that have abstracts. The title, abstract and keywords of each entry constituted the “bag of words” to be analyzed for each individual document.

Table 1. The number of documents from each publication venue in the information visualization dataset.

Source name	count
Proceedings of the IEEE Symposium on Information Visualization	152
IEEE Visualization	32
Lecture Notes In Computer Science (LNCS)	22
Conference on Human Factors in Computing Systems (SIGCHI)	21
ACM CSUR and Transactions (TOCS,TOID,TOG)	21
Symposium on User Interface Software and Technology (UIST)	17
IEEE Computer Graphics and Applications	13
IEEE Transactions	13
International Conference on Computer Graphics and Interactive Techniques (CGIT)	12
Communications of the ACM (CACM)	10
Advanced Visual Interfaces (AVI)	10
Other	106

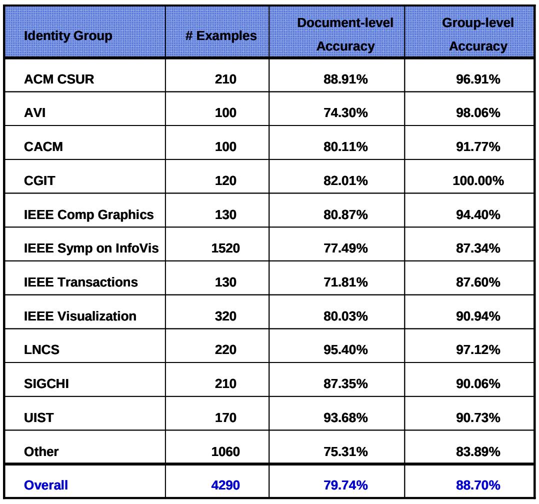
Several publication venues were represented here. We arbitrarily decided to consider each venue having 10 or more publications in the dataset as an identity group. This requirement provided some assurance that enough information was available to determine useful topic profiles for each group. Papers that did not belong to a venue meeting this minimum requirement were lumped together into a default group called “Other”. Table 1 lists the identity groups and the number of documents associated with them. Given that the field of information visualization is itself a specialized identity group, it is not obvious that the smaller groups we specify here will have any distinguishable properties. There is a strong topic interdependency among these groups, which makes for a challenging classification task. It is also not obvious that broadly inclusive groups like the IEEE Symposium on Information Visualization or the “Other” category will have any distinguishable properties since they include papers from all the relevant topic areas in the field.

The selected documents were preprocessed to convert all words to lower case and remove all punctuation, single characters, and two-character words. We also excluded words on a standard “stop” list of words used in computational linguistics (e.g., numbers, function words like “the” and “of”, etc.) and words that appeared fewer than five times in the corpus. Words were not stemmed, except to remove any trailing “s” after a consonant. This preprocessing resulted in a vocabulary of 1405 unique words and a total of 31,256 word tokens in the selected documents.

Two sets of topic features were computed from this collection: one from the set of individual documents and one from the set of aggregated mega-documents associated with each group. The optimal number of topics that fits the data well without overfitting was determined using a Bayesian method for solving model selection problems [4]. The model with 100 topics had the highest likelihood for the collection of mega-documents and the 200 topic model was best for the collection of individual documents. The resulting feature vector descriptions of each document were then used as examples for training classifiers that discriminate one identity group from another. Examples were generated by running the LDA parameter estimation algorithm over the words in a document for 500 iterations and drawing a sample of the topic distribution at the end of the run. We used a sparse representation for each example, only listing features which had a weight greater than the weight for a random choice. Ten examples were generated for each document[1], producing an overall total of 4290 examples for each feature set.

On each of ten runs of the experiment, a support vector machine classifier [6] was trained with a random subset of 75% of the examples, with the remaining 25% used for testing[2]. For each group, we trained a "one-versus-the-rest" binary classifier. This means that the examples from one class became positive examples while the examples from all other classes were treated as negative examples in a binary classification task. The overall solution to the multi-class problem is given by following the recommendation of the binary classifier with the highest classification score on a given test example. The support vector machine used a radial basis function kernel along with a cost factor to compensate for the unbalanced number of positive and negative examples. The cost factor weighs classification errors so that the total cost of false positives is roughly equal to the total cost of false negatives. For details about the cost factor, see [8].

Table 2. Results of the classification experiments on the information visualization dataset.



Identity Group	# Examples	Document-level Accuracy	Group-level Accuracy
ACM CSUR	210	88.91%	96.91%
AVI	100	74.30%	98.06%
CACM	100	80.11%	91.77%
CGIT	120	82.01%	100.00%
IEEE Comp Graphics	130	80.87%	94.40%
IEEE Symp on InfoVis	1520	77.49%	87.34%
IEEE Transactions	130	71.81%	87.60%
IEEE Visualization	320	80.03%	90.94%
LNCS	220	95.40%	97.12%
SIGCHI	210	87.35%	90.06%
UIST	170	93.68%	90.73%
Other	1060	75.31%	83.89%
Overall	4290	79.74%	88.70%

[1] Since the algorithm and representation are probabilistic, different runs will produce different feature vectors. The "true" feature vector can be thought of as a prototype that is most representative of all possible sample vectors.

[2] More specifically, we used 10-fold cross validation with each fold independently chosen, a method sometimes referred to as random subsampling.

The results of the document classification experiments are summarized in Table 2. Classification accuracy was computed by averaging over the ten independent runs. These results show that the group-level features produce a statistically significant improvement in overall classification accuracy over the document-level features on test data (88.7% accuracy versus 79.7% accuracy). Not surprisingly, the two most diverse classes ("IEEE Symposium on Information Visualization" and "Other") had the worst classification performance. The confusion matrix data shows that most classification errors were due to erroneous assignments to one of these classes.

These empirical results suggest that group-level attributes can provide better predictions of the contents of scientific documents published by group members than predictions based on attributes derived from the individual documents. This approach to document classification represents a modest step toward developing new approaches to modeling the effects of social identity groups on the behavior of individual members.

3 Authorship Analysis

Though the document classification approach based on topic analysis was designed to extract social identity group "fingerprints" from a document collection, it can be applied to a wide variety of document classification problems. The topic analysis does not depend on the existence of a social network of people who interact with each other[3]. The only requirement is that group or class labels are available for the document collection used for training.

In this section we show how our approach to document classification can be used in a setting that is relevant to forensic authorship analysis. We revisit a study [9] of how age and gender differences among bloggers are reflected in the writing style and content of blogs.

3.1 Age and Gender Effects on Blog Data

The Blog Authorship Corpus [9] was constructed using blogs collected from blogger.com in August 2004. Each blog selected for the corpus contained at least 500 words, including at least 200 occurrences of common English words, along with author-provided information about gender and age. From an initial collection of 46,947 blogs, a subset was extracted that included bloggers in three age categories: "10s" (ages 13-17), "20s" (ages 23-27), and "30+" (ages 33-47). Blogs in the "boundary" age groups 18-22 and 28-32 were removed in order to facilitate more reliable age categorization. Within each age category, the gender distribution was equalized by randomly discarding excess blogs from the larger gender group, leaving 8,240 "10s" blogs, 8,086 "20s" blogs and 2,994 "30+" blogs. The final corpus consists of 19,320 blogs containing 681,288 posts and over 140 million words (yielding approximately 35 posts and 7250 words per blog).

[3] Moreover, the attributes shared by the group do not need to be linked to social identity.

Schler *et al*. [9] used this corpus in a study that characterized differences in blogs based on style-related features and content-related features. Three types of style related features were considered: parts of speech (auxiliary verbs, pronouns, etc.); function words (frequently occurring English words such as “a”, “it”, “the”, “very”, etc.); and, “blogs words” – such as lol - and hyperlinks. The content-related features were simple content words that had the highest information gain among the frequently appearing words in a category. These content words were often closely associated with some distinct theme or topic. For example, the words “mother”, “father”, and “kids” are related to the theme “family”.

In order to show that these vocabulary features could be used to predict the age and gender of a blog’s author, Schler *et al*. constructed classification models for automated author profiling. Each blog was represented by a numeric vector whose entries were the frequencies, in that blog, of 502 style-related feature and 1000 content-related features. A linear-threshold machine learning algorithm [7] was applied to these vectors to generate classification models for author age and author gender. Empirical results show that these models can automatically classify unseen documents into the correct age category with an accuracy of 76.2% and identify the correct gender with an accuracy of 80.1%. This suggests that the vocabulary features apparently do capture important differences in writing style and content that distinguish bloggers with different genders and in different age categories.

3.2 Using Topic Analysis to Compute Age and Gender Attributes

Instead of trying to carefully select a subset of words as features that will discriminate between various age and gender categories, an intriguing alternative is to make automated feature discovery part of the authorship analysis problem. Features extracted using topic analysis techniques reflect a coupling between style and content as indicated by word co-occurrence patterns. This synergy may produce classification models with performance that is comparable to what can be obtained using hand-selected vocabulary features.

In order to test this hypothesis, we applied our topic analysis methodology to the Blog Authorship Corpus. We followed the same procedure used for the Information Visualization dataset, generating a model of 100 topics for the age group-level attributes and a separate model of 100 topics for the gender group-level attributes. These models were derived from a vocabulary of 148,201 unique words and a total of 26,048,869 word tokens in the corpus. Results show that these models can automatically classify unseen documents into the correct age category with an accuracy of 72.83% and identify the correct gender with an accuracy of 75.04%. This compares favorably with the results reported by Schler *et al*. that were based on a much larger number of hand-selected features.

4 Summary

This paper has shown how identity group "fingerprints" can be extracted from a document collection by applying topic analysis methods in a novel way. Empirical results on scientific publication data suggest that group-level attributes can provide better predictions of the contents of scientific documents published by group members than predictions based on attributes derived from the individual documents. This argues in favor of using group-level attributes for document classification.

Experiments with blog data show that this document classification method can also be effective for forensic authorship analysis tasks. Besides providing good classification accuracy, it has the added benefit of automatically inferring a good set of features that appear to account for both style-related and content-related differences in vocabulary usage. This capability could be a useful supplement to forensic methods that incorporate other techniques such as linguistics and behavioral profiling.

Acknowledgments. This research was funded by the MITRE Sponsored Research program. That support is gratefully acknowledged.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Börner, K., Dall'asta, L., Ke, W., Vespignani, A.: Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. Complexity 10(4), 57–67 (2005)
3. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-mail Content for Author Identification Forensics. SIMOD Record, 30(4), 55–64 (2001)
4. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Science 101, 5228–5235 (2004)
5. Griffiths, T., Steyvers, M., Tenenbaum, J.: Topics in semantic representation. Psychological Review 114(2), 211–244 (2007)
6. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge (1999)
7. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linear–threshold algorithm. Machine Learning 2(4), 285–318 (1988)
8. Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledge-based approach – A case study in intensive care monitoring. In: Proceedings of the 16th International Conference on Machine Learning, pp. 268–277. Morgan Kaufmann, San Francisco (1999)
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium. AAAI Press, Menlo Park (2006)
10. Stokar von Neuforn, D., Franke, K.: Reading Between the Lines: Human-centered Classification of Communication Patterns and Intentions. In: Liu, H., Salerno, J., Young, M. (eds.) Social Computing, Behavioral Modeling, and Prediction. Springer, New York (2008)