Authorship Attribution with Topic Drift Model

Min Yang,[1,2] **Dingju Zhu,**[1*] **Yong Tang,**[1] **Jingxuan Wang**[2]
[1]South China Normal University, China
[2]University of Hong Kong, Hong Kong
myang@cs.hku.hk {zhudj,ytang}@scnu.edu.cn jingxuan@hku.hk

Abstract

Detecting authorship attribution is an active research direction due to its legal and financial importance. The goal is to identify the authorship of anonymous texts. In this paper, we propose a Topic Drift Model (TDM), monitoring the dynamicity of authors' writing style and latent topics of interest. Our model is sensitive to the temporal information and the ordering of words, thus it extracts more information from texts.

Introduction

Most of the early work on AA focuses on formal texts with only a few candidate authors while researchers have recently turned their attention to informal texts and tens of thousands of authors (see Section 2). As the length of texts decreases and the number of candidate authors increases, finding the author of texts remains a challenge.

Capturing authors' writing style is crucial for authorship attribution. Most current AA approaches neglect the temporal changes in authors' writing style, which may lead to poor performance of AA for authors who change their writing styles constantly. Moreover, most existing AA models make the bag-of-word assumption, which neglects the ordering of words and semantics of the context though these factors are important in identifying author's writing.

Inspired by (Yang, Cui, and Tu 2015; Yang et al. 2016), in this paper, we introduce a novel Topic Drift Model (TDM) for modeling the dynamic evolution of individual author's latent topics of interest. TDM learns the representation of words and authors' latent topics as vectors. The similarity between these vectors may be represented by their Euclidean distance. Each author's topics of interest are represented by a sequence of vectors, where the speed of the topic drifting is controlled by the similarity of consecutive vectors. The TDM model also captures the fact that co-authors usually share common topics, by making their interest vectors positively correlated. Finally, we assign the given text to the author whose interest vector has highest similarity with the author interest vector of the given text.

Our model has three advantages over the previous similarity based AA approaches. First, it captures the fact that the authors' interests and writing styles may change over time, and automatically learn this drift from raw data. Second, our model provides a natural way to measure the similarity of latent topics of interest, which helps modeling the speed of interest drift, as well as measuring the similarity between the author and the text. Third, our model takes the ordering of words into account, so that the topics reflect the semantics of the context.

Model description

We apply a Topic Drift Model (TDM) (Yang et al. 2016) for modeling the dynamic evolution of individual author's interest. We assume that there are W different words in the vocabulary and there are D documents in corpus. In addition, these documents belong to T topics, where T is a hyperparameter specified by the user. We use an p-dimensional interest vector $\mathrm{vec}(a, d) \in \mathbb{R}^p$ to represent the interests of author a when he/she writes document d. Two authors have similar interests if the distance between their interest vectors is small.

Let t be the time that document d was written. Let d' be the last document that author a has written before he/she writes the current document, and let the timestamps for d' be t' ($t' \leq t$). We define a joint distribution on the vectors $\mathrm{vec}(a, d)$. It is a multivariate normal distribution on $\mathbb{R}^p$ taking the form $N(\mu_d, \Sigma_d)$. Firstly, we specify the mean $\mu_d := vec(a{,}d')$. This definition implies that the new interests of authors have connections to the history. Second, we define the covariance matrix to characterize the interest drift of the author $\Sigma_d := \sigma(t - t')\sigma(t - t')I$. Here, $\sigma(x)$ is an increasing function of x. It means that as more time passed, the covariance matrix entries get bigger, indicating that the interest of author is less likely to concentrate on its mean – the interest vector when he/she wrote the earlier document d'.

Following the idea of (Yang, Cui, and Tu 2015), we represent the topic model as Gaussian mixture model of vectors which encode words, sentences and documents. Each mixture component is associated with a specific topic. Given the Gaussian mixture model λ, the generative process is described as follow: for each word w in the vocabulary, we sample its topic $z(w)$ from the multinomial distribution $\pi := (\pi_1, \pi_2, \ldots, \pi_T)$ (T is the number of topics) and sample its vector representation $\mathrm{vec}(w)$ from Gaussian distribu-

[*] Corresponding author

tion $\mathcal{N}(\mu_{z(w)}, \Sigma_{z(w)})$. For each document d and each sentence s in the document, we sample their topics $z(d)$, $z(s)$ from distribution π and sample their vector representations, namely $\text{vec}(d)$ and $\text{vec}(s)$, also from the Gaussian mixture model.

By estimating the model parameters, we learn the word representations that make one word predictable from its previous words, the context and its author interest vector. Jointly, we learn the distribution of topics that words, sentences and documents belong to. The parameters of the model are learnt through a prediction task: given vectors associated with the context, the goal is to predict each word in the document. We refer the reader to (Yang, Cui, and Tu 2015) for a more detailed implementation.

Authorship attribution

Given our Topic Drift Model, we assume that a new text d_{new} was written by an author a_{new} at time t_{new}. Since the author is unknown, we treat her as a new author, then use the model to infer the posterior distribution of the author's interest (i.e., vector $\text{vec}(a_{new}, d_{new})$). We then calculate the similarity between every candidate author's interest vector with the "new" author's vector. The most similar candidate author is returned as the writer of the document.

Since each author's topics of interest are represented by a sequence of vectors $\{\text{vec}(a, d_i) | i = 1, ...D\}$, we use Nadaraya-Watson kernel regression algorithm (Nadaraya 1964) to estimate each author's interest vector at t_{new} as a locally weighted average:

$$\tilde{\text{vec}}(a, d_{new}) = \frac{\sum_{i=1}^{D} K(t_{new}, t_i) * \text{vec}(a, d_i)}{\sum_{i=1}^{T} K(t_{new}, t_i)} \qquad (1)$$

where D is the total number of documents written by author a, t_i is the time when a wrote document d_i, and $K(\cdot, \cdot)$ can be the RBF kernel.

Finally, we calculate the similarity between $\text{vec}(a_{new}, d_{new})$ and $\tilde{\text{vec}}(a, d_{new})$ by the Euclidean distance. We assign the given text to the author whose interest vector has highest similarity with the "new" author of the given text.

Experiments

Datasets

PAN'11 emails (PAN'11): This corpus contains 9337 documents by 72 different authors [1]. We ran the method on the corresponding testing set that only contains authors in the training set.

Blog: This corpus is the largest dataset that is widely used for authorship attribution, containing 678,161 blog posts by 19,320 authors from blogger.com in August 2004 (Schler et al. 2006).

Baseline methods

We compare our approach with several state of the art baseline methods, including SVM, LDA-H (Seroussi, Zukerman, and Bohnert 2011), DADT (Seroussi, Bohnert, and Zukerman 2012), TFS (Azarbonyad et al. 2015) and NNLM (Ge, Sun, and Smith 2016).

[1] http://www.cs.cmu.edu/˜enron/

Accuracy	SVM	LDA-H	DADT	TFS	NNLM	TDM
PAN'11	0.502	0.480	0.514	0.510	0.482	0.542
Blog	0.246	0.08	0.280	0.270	0.252	0.308

Table 1: The accuracies on PAN'11 and Blog datasets

Experimental results

We first vary the number of latent topics of the models to see how the performance changes. We conduct experiments with 5, 10, 20, 40, 60, 80, 100, 150, 200 topics. Due to space limitations, we refer the reader to our complementary-material webpage for a more detailed experimental results. Table 1 shows the best accuracies of all the models. The best accuracy of our model is consistently and clearly better than that of other models on the three data sets. For example, for PAN'11 data set, the best accuracy of DADT and TFS are 51.6% and 51.2% respectively, which are slightly higher than SVM, LDA-H and NNLM. Our model further improves the accuracy to 53.9%.

Conclusion

In this paper, we have proposed a Topic drift model (TDM) for authorship attribution, which explicitly characterizes the dynamic topics of interest drifting for individual authors.

Acknowledgements Major Project of National Social Science Fund (grant No. is 14ZDB101).

References

Azarbonyad, H.; Dehghani, M.; Marx, M.; and Kamps, J. 2015. Time-aware authorship attribution for short text streams. In *ACM SIGIR*, 727–730.

Ge, Z.; Sun, Y.; and Smith, M. J. 2016. Authorship attribution using a neural network language model. *arXiv preprint arXiv:1602.05292*.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9(1):141–142.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. W. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium*, volume 6, 199–205.

Seroussi, Y.; Bohnert, F.; and Zukerman, I. 2012. Authorship attribution with author-aware topic models. In *ACL*, 264–269.

Seroussi, Y.; Zukerman, I.; and Bohnert, F. 2011. Authorship attribution with latent dirichlet allocation. In *CoNLL*, 181–189.

Yang, M.; Mei, J.; Xu, F.; Tu, W.; and Lu, Z. 2016. Discovering author interest evolution in topic modeling. In *ACM SIGIR*, 801–804.

Yang, M.; Cui, T.; and Tu, W. 2015. Ordering-sensitive and semantic-aware topic modeling. In *AAAI*, 2353–2359.