Time-Aware Authorship Attribution for Short Text Streams

Hosein Azarbonyad
Informatics Institute, University of Amsterdam
h.azarbonyad@uva.nl

Mostafa Dehghani
Institute for Logic, Language and Computation, University of Amsterdam
dehghani@uva.nl

Maarten Marx
Informatics Institute, University of Amsterdam
maartenmarx@uva.nl

Jaap Kamps
Institute for Logic, Language and Computation, University of Amsterdam
kamps@uva.nl

ABSTRACT

Identifying authors of short texts on Internet or social media based communication systems is an important tool against fraud and cybercrimes. Besides the challenges raised by the limited length of these short messages, evolving language and writing styles of authors of these texts makes authorship attribution difficult. Most current short text authorship attribution approaches only address the challenge of limited text length. However, neglecting the second challenge may lead to poor performance of authorship attribution for authors who change their writing styles.

In this paper, we analyse the temporal changes of word usage by authors of tweets and emails and based on this analysis we propose an approach to estimate the dynamicity of authors' word usage. The proposed approach is inspired by time-aware language models and can be employed in any time-unaware authorship attribution method. Our experiments on Tweets and the Enron email dataset show that the proposed time-aware authorship attribution approach significantly outperforms baselines that neglect the dynamicity of authors.

Categories and Subject Descriptors

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

Keywords

Authorship Attribution; Time-Aware Language Models; Short Text Analysis

1. INTRODUCTION

Automatic authorship attribution is a growing research direction due to its legal and financial importance [12]. In the recent decade with the growth of Internet based communication facilities, much content on the web is in the form of short messages. Finding the author of a short message is important since much fraud and cybercrimes occur with exchanging emails and short messages. Usually cybercriminals tend to use an anonymous identity in the Internet based communication systems. Therefore, finding the authors of short texts could be useful for law enforcement agencies. As length of texts decreases, finding the author of the texts becomes more challenging [8, 10, 11].

Current authorship attribution approaches neglect an important factor in human development: as a person matures or a significant event occurs in his life (such as changing job, getting married, moving in a new circle of friends, etc) over time the model of his writing style and the words used may change as well. As an example, Lancashire et al [7] analyses the temporal changes on the vocabulary usage by Agatha Christie and conclude that her vocabulary-size decreases over time. Cf. also [13]. As another example, Figure 1 shows the temporal changes of vocabulary usages of 133 Twitter users over a period of 40 months. The figure shows that the similarity of content to a fixed static corpus decreases over time. In fact, we can conclude that content generated at the current time is more similar to recent content than to older content. Current authorship attribution approaches neglect this fact and use all material generated by authors with the same influence.

This paper tries to answer two crucial questions in authorship attribution for short texts. The first research question is: Does the writing style of authors of short text change over time? And if so, do they change their writing styles by the same rate? The second research question is: How does the temporal change of writing styles of authors affect authorship attribution? And how we can capture the changes in the writing styles of authors and take the changes into account to overcome the effects of drift in authorship attribution?

We answer these questions using two datasets: one is collected from Twitter and the Enron email corpus [3]. We introduce a new time-aware authorship attribution approach which is inspired by time-based language models [9] and can be employed in any time-unaware authorship attribution method to consider the temporal drifts in authorship attribution process. Our evaluations on tweets and Enron datasets show that the proposed time-aware approach is able to incorporate the temporal changes in authors writing styles and outperforms two competitive baselines.

The paper is organized as follows. We review related work in Section 2. Section 3 contains our time-aware authorship attribution framework. The datasets and experiments are described in Section 4. We conclude in Section 5.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767799 .

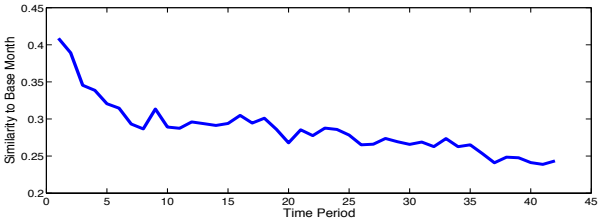

Figure 1: Vocabulary usage changes of Twitter users over time. A dataset containing 133 Twitter users and their written tweets is collected. The first two months of the users' activity in Twitter are considered as start period. Also, each following month is considered as a time period. The $x-axis$ shows the time periods and $y-axis$ shows the averaged similarity of the contents generated by the users at each time period with the content generated by them in the base time period. Cosine similarity over frequency of character 4-grams in users' contents is employed as similarity measure.

2. RELATED WORK

Authorship attribution approaches can be categorized into two main categories: similarity based approaches and machine learning based approaches [12]. Several studies showed that similarity based approaches outperform machine learning based methods when the number of candidate authors is high [5, 8]. In this paper, we only use similarity based approaches. Previous work showed that character n-grams are the most effective units for calculating the similarity of a given text with authors' profiles [1, 4, 12]. The SCAP method[1], which simply calculates the Jacard similarity of a given text and the profile texts of authors and assigns the given text to the most similar author, is the simplest similarity based approach. Among different similarity based methods, the *feature sampling method* proposed in [4] is the best performing method for authorship attribution. This method samples from all features (i.e., character n-grams) and calculates the similarity of the given text with all authors' profiles on the sampled features using cosine similarity. This process is repeated k times and the text is assigned to the author whose profile is most similar to the given text for a certain fixed number of the k times. (If no author achieves this threshold, no one is assigned). A variation of this method outperformed other approaches in the Authorship Verification task in which the goal is to determine if two documents are written by the same author [6]. Recent research focuses on authorship attribution for short texts [11, 8, 10]. Incorporating temporal changes of writing styles of authors has not been applied before in authorship attribution.

3. TIME-AWARE AUTHORSHIP ATTRIBUTION

Time-based language models have been shown to be effective in temporal information retrieval [9]. We use a similar approach in authorship attribution when we consider the temporal changes in authors' writings. We first divide the whole timeline of an author in time periods of a fixed length and then construct a language model for each period. In this work, we use character n-grams for our language models. In fact, the language model of each period is a probability distribution over n-grams of the texts generated in that period. For a new generated short text, we calculate its similarity with the language model constructed for each period weighted by a decay factor which is a function of the temporal difference of the date of the short text with the period. More specifically, the time-aware probability that a given short text s is written by an author a is calculated as follows:

$$P(s|a) = \sum_{t \in T \wedge t < t_s} decay(t_s - t) * P(s|\theta_{a_t}), \quad (1)$$

where T is set of all periods. We discretize the whole timeline to T periods. $P(s|\theta_{a_t})$ is the probability that s is generated by the language model of author a in time period t. The function $decay()$ is a monotonically decreasing function, giving less weight to older periods. In section 3.1, we introduce different decay functions and study their effectiveness. We use character 4-grams as token units (features) and employ unigram language models [14] for estimating $P(s|\theta_{a_t})$. We construct a language model for t and first estimeate the likelihood of generating text s from language model of author a in time t as follows:

$$P(s|\theta_{a_t}) = \prod_{ng \in ngrams_s} p(ng|\theta_{a_t})^{c(ng, ngrams_s)}, \quad (2)$$

where θ_{a_t} is the language model of a in time period t, $ngrams_s$ is all character n-grams extracted from s, and $c(ng, ngrams_s)$ is the frequency of n-gram ng in $ngrams_s$. $p(ng|\theta_{a_t})$ is calculated as follows using Jelinek-Mercer smoothing:

$$p(ng|\theta_{a_t}) = (1-\lambda) * p_{ml}(ng|d_{a_t}) + \lambda * p(ng|\mathcal{C}), \quad (3)$$

where, d_{a_t} is a document containing all character n-grams of texts generated by a in time period t, $p_{ml}(ng|d_{a_t})$ is estimated using maximum likelihood, $\mathcal{C}$ is all character n-grams of whole corpus and $p(ng|\mathcal{C})$ is also estimated using maximum likelihood.

It is supposed that every author is equally likely before any piece of text is given and finally, the author of s is determined as follows:

$$\hat{a} = \underset{a}{\text{argmax}}\ P(s|a) \quad (4)$$

We assign s to $\hat{a}$ if $P(s|a)$ is more than a predefined threshold.

We use this approach to extend the SCAP method[1] and the feature sampling method [4]. We use (1) to calculate similarity. When $decay()$ is the constant function assigning 1, we have a time-unaware approach. Otherwise the appraoch is time aware.

3.1 Decay Factor

Several decay functions have been proposed [2, 9]. In this paper, we compare a general decay function which is same for all users to a function which is specific for each user. The exponential decaying function is the most used method in temporal IR. However, Figure 1 shows that the writing styles of authors do not change that dramatically. Therefore, we use a linear decreasing function. We employ the slope of the curve plotted in Figure 1 as the slope of the general decay function. We use linear regression to estimate the slope and intercept parameters. Finally, the decay factor for a time period t is then calculated as follows:

$$decay(t) = \frac{1}{Z}(at + b), \quad (5)$$

where Z is the normalizing factor (the sum of decay values for all periods should be 1), a and b are the parameters of the linear degrading function.

Changes in writing styles need not be the same for all authors. For estimating the parameters of specific degrading function for an author a, we first calculate the similarity of contents generated by a in each time period with the contents generated by him/her in start period and then plot a vocabulary usage change curve for a (similar to Figure 1). Then we use linear regression to estimate

the slope and intercept parametrs of the plotted curve and employ the estimated parameters as the parameters of the specific decay function. Finally, the decay function for an author a at a time period t is calculated as follows:

$$decay_a(t) = \frac{1}{Z_a}(a_a t + b_a), \quad (6)$$

where a_a and b_a are the parameters of decay function for the author a.

4. DATASETS AND EXPERIMENTAL RESULTS

We describe the used datasets, explain the experiments and report our results.

4.1 Datasets

We use two different datasets: tweets and the Enron email corpus [3]. Both datasets contain short messages generated over a long period, and thus they are suitable for time-aware authorship attribution of short texts.

We collected tweets from users of Twitter who have tweeted for a long time. This dataset contains 133 users. The average number of tweets per user is 1820 and the average number of tweets of each user per month is 31. Tweets are written between 2010-01 and 2014-10. We divided this period into 46 months. Since character 4-grams have been shown to be the most effective units in authorship attribution [6, 8], we model texts and authors using unigram language models consisting of character 4-grams. The average and median number of 4-grams per month in this dataset are 61,677 and 101,867 respectively. The average length of tweets in this dataset is 101 characters with a the standard deviation of 42.

From the Enron dataset we selected mails from 15 prolific authors written between 1998-01 and 2002-09. We divided the corpus into 45 months. The average number of emails per person in this dataset is 3200 and the average and median number of emails written by a user per month in the selected dataset are 68 and 74 respectively. The average and median number of 4-grams per month in this dataset are 436,940 and 478,678 respectively. The average length of emails in this dataset is 648 characters with a standard deviation of 1253 and the median of 606.

4.2 Experimental results

We now discuss our results, following the two main research questions.

4.2.1 *Drift in word usage over time*

Starting with our first research question: Does the writing style of authors of short text change over time? And if so, do they change their writing styles by the same rate? Figure 1 shows that on average there is a change in vocabulary usages of authors. However, we expect that different users have different vocabulary usage change rates. Table 1 shows the statistics of the slopes of specific decay functions. We use the method described in Section 3.1 for estimating the slope of vocabulary usage change plot for each user. For calculating the similarity we use frequency of 4-grams. As can be seen from this table, the average of slopes of change plots of all users is almost zero. However there is a relatively large difference between maximum and minimum values of slopes in both datasets. This indicates that different authors change their vocabulary usage in different rates over the time.

Dataset	max	min	$average$	$median$	std
Enron(N=15)	0.09	-0.1	-0.007	-0.006	0.161
Tweets(N=133)	0.06	-0.07	-0.001	-0.002	0.07

Table 1: The statistics of slope a_a in the decay function 6 for different authors on Enron and Tweet dataset.

Dataset	Time-unaware	Time-aware(general)	Time-aware(specific)
Enron	0.87	0.88	0.94 (8%)$^{\blacktriangle}$
Tweets	0.69	0.71	0.80 (15%)$^{\blacktriangle}$

Table 2: Precision of feature sampling method on tweets and Enron dataset at Recall point of 0.3. ($^{\blacktriangle}$ indicates the significance using t-test, one-tailed, $p-value < 0.05$)

4.2.2 *Results of time-aware authorship attribution methods*

To see the effect of drift in authorship attribution, we first do time-unaware authorship attribution. We use 90% oldest messages generated by each author for constructing the author-model and use the remaining 10% newest messages to test different authorship attribution approaches. In time-aware approaches, we tested different lengths for time periods. Based on our experiments the best performance on both tweets and Enron datasets is achieved when we set the length of time periods to one month.

Figure 2 shows the precision-recall curves of time-aware SCAP, time-unaware SCAP, time-aware feature sampling, and time-unaware feature sampling methods on tweets and Enron datasets. We assign the given text to the found author if the similarity of the text with the author's model is more than a predefined threshold otherwise no authorship is made for the text. By changing the value of the threshold we achieved different values of precision and recall and plotted them in Figure 2. Precision is the proportion of correct attributions among all attributions made by the method and recall is the proportion of test samples for which and attribution made by the method and is correcet. As decay function we use specific decay function (Equation 6). We tested different values for the parameter of Jelinek-Mercer smoothing method λ and the number of iterations of feature sampling method k. The best results are achieved with $\lambda = 0.3$ and $k = 100$. In the results, these parameter values are used. Figure 2 shows that the precision of the time-aware approach is higher than the precision of time-unaware approach at different recall points both when we apply time-aware approach on simple SCAP method and when we apply it on feature sampling method. Also, from the results it can be concluded that the time-aware feature sampling method is the best performing method on both datasets.

The precision of feature sampling method using general and specific decay functions and without using decay function at recall of 0.3 is shown in Table 2. The results show that using the specific decay function gives significantly better performance compared to using the general decay function. Also, the precision of the time-aware method which uses general decay function is almost same as the precision of the time-unaware approach. In fact using a general decay function does not help to track the changes of writing styles of auhtors and can not distinguish between dynamic and static authors. The high value of precison on Enron dataset is mainly due to the low number of authors in this dataset which makes attribution easier.

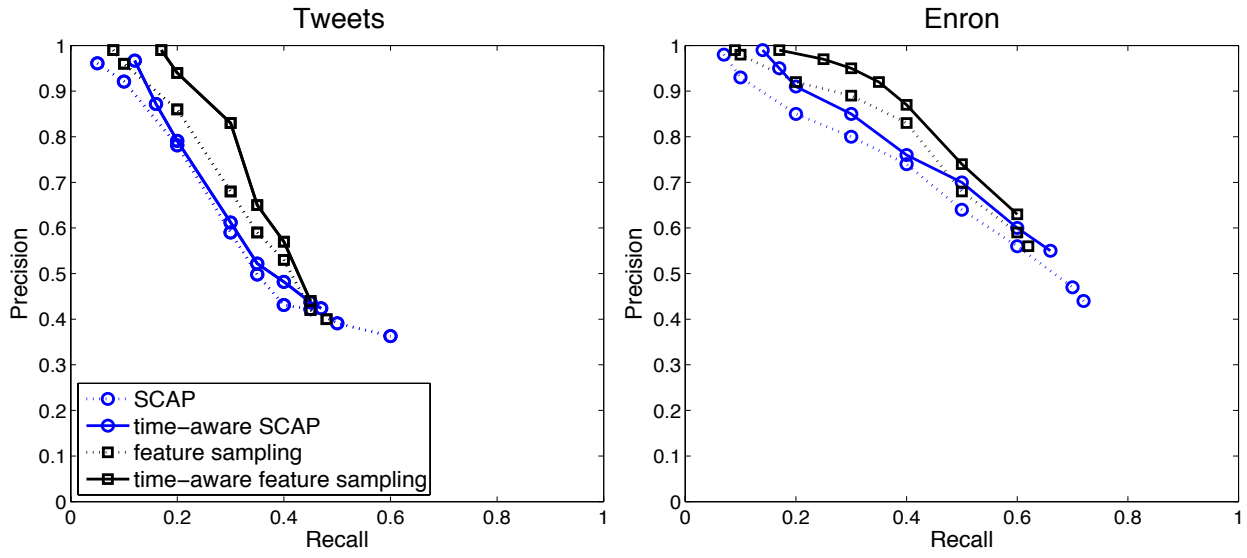

Figure 2: Precison-Recall curves of time-aware and time-unaware authorship attribution methods on tweets and Enron datasets.

5. CONCLUSIONS

We considered the effect of temporal change in the writing style and vocabulary usage of authors on the task of authorship attribution for short text streams. Our findings were based on two datasets, the Enron email corpus and an own-generated set of tweets from 133 authors who tweeted over a period of almost 4 years. We first investigated whether temporal change is a problem at all, and found that authors do change and that different authors change differently. We used a linearly decreasing temporal decay function to incorporate the temporal changes of authors' vocabulary usage in authorship attribution. We divided the whole timeline of authors into fixed size periods and constructed a language model for each period. In the evaluation we created time aware versions of two commonly used authorship attribution methods. In both methods the time-aware version performed significantly better than the "static" version, over both datasets.

Feature engineering is an interesting future research direction. Some (types of) features may change more than others, and we could incorporate that in our models.

Acknowledgements This research was supported by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014) and by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement ENVRI, number 283465.

6. REFERENCES

[1] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski, and B. S. Howald. Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. *IJDE*, 6(1), 2007.

[2] N. Kanhabua and K. Nørvåg. A comparison of time-aware ranking methods. SIGIR '11, pages 1257–1258, 2011.

[3] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *ECML'04*, pages 217–226, 2004.

[4] M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. *LREC*, 45(1):83–94, 2011.

[5] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. SIGIR '06, pages 659–660, 2006.

[6] M. Koppel and Y. Winter. Determining if two documents are written by the same author. *JASIST*, 65(1):178–187, 2014.

[7] I. Lancashire and G. Hirst. Vocabulary changes in agatha christie's mysteries as an indication of dementia: A case study. In *19th Annual Rotman Research Institute Conference Cognitive Aging: Research and Practice*, pages 1–5, 2009.

[8] R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC)*, pages 1–8, 2010.

[9] X. Li and W. B. Croft. Time-based language models. CIKM '03, pages 469–475, 2003.

[10] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship attribution of micro-messages. In *EMNLP'13*, pages 1880–1891, 2013.

[11] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, and B. Maia. Twazn me: Automatic authorship analysis of micro-blogging messages. NLDB'11, pages 161–168, 2011.

[12] E. Stamatatos. A survey of modern authorship attribution methods. *JASIST*, 60(3):538–556, 2009.

[13] M. van Dam and C. Hauff. Large-scale author verification: Temporal and topical influences. SIGIR '14, pages 1039–1042, 2014.

[14] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, 22(2):179–214, 2004.