Time-Aware Authorship Attribution of Short Texts

Extended Abstract

Hosein Azarbonyad[1] Mostafa Dehghani[2] Maarten Marx[1] Jaap Kamps[2]

[1]Informatics Institute, University of Amsterdam, The Netherlands
[2]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
{h.azarbonyad,dehghani,maartenmarx,kamps}@uva.nl

ABSTRACT

Automatic authorship attribution is a growing research direction due to its legal and financial importance. 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.

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. 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 [5] and can be employed in any time-unaware authorship attribution method to consider the temporal drifts in authorship attribution process. 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. 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. The time-aware probability that a given short text s is written by an author a is calculated as follows:

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

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. We estimate two different decay functions: a general decay function which is same for all authors and a specific decay function for each author estimated based on the change rates of writing styles of authors.

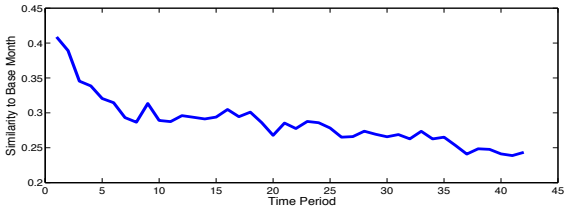

Figure 1: Vocabulary usage changes of Twitter users over time. A dataset containing 133 Twitter users and their written tweets is collected. The first two months of the users' activity in Twitter are considered as start period. Also, each following month is considered as a time period. The $x-axis$ shows 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.

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

$$\hat{a} = \underset{a}{\operatorname{argmax}}\ P(s|a) \qquad (2)$$

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[2] and the feature sampling method [4]. We consider SCAP and feature sampling methods as our baselines. 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 proposed time-aware method improves the accuracy of time-unaware feature sampling baseline by 8% on Enron dataset and by 15% on Tweets dataset. Also this method improves the accuracy of SCAP method by 17% on Enron dataset and by 27% on Tweets dataset.

Acknowledgements The full version of this paper is available as [1]. 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.

References

[1] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps. Time-aware authorship attribution for short text streams. SIGIR '15, pages 727–730, 2015.

[2] 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.

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