Authorship attribution for short texts with Top-N Position-Frequency feature set

No Author Given

No Institute Given

Abstract. In this paper, we propose a novel method Top-N Position-Frequency (named as TNPF) that measures the importance of features appropriately for authorship attribution on short texts. We study the effects of different features measurements and the scale of the feature set for authorship attribution. Two real datasets Enron email and Amazon food review are applied for our experiments. We empirically show that TNPF feature set achieves better classification performance than the feature sets with traditional frequency measurements.

1 Introduction

Authorship attribution has developed rapidly over the past decades due to its legal and financial importance[13]. It helps people to determine the author of the controversial texts. In the recent years, with the development of social media, cybercrime has become very rampant on the Internet. To prevent it, finding the author of short texts becomes very necessary, for the cybercrime usually occurs with exchanging short texts like emails[2]. However, short texts typically have few words, unclear structure, and irregular usage compared with long texts. The restrictions make authorship attribution difficult, and the methods that work well on long texts cannot get the same performance on short texts[9]. As a result, many researchers begin to solve the problem of authorship attribution for short texts, such as emails[1] and blogs[7].

Authorship attribution methods can be divided into two main categories: similarity-based methods and machine-learning-based methods. When it comes to short texts, like emails or twitters, the similarity-based methods usually perform better than the machine learning based methods because of the large number of candidate authors[9]. Our paper also conducts researches on the similarity-based methods.

Almost all current researchers focus on how to build a new kind of feature or an efficient classification algorithm to improve the performance of classification. Some researchers set some limitations on feature selection and want to define a new kind of representative feature to reflect author's style, like k-signature[11] or time[2]. There are also many useful algorithms in classification like SCAP[5] and feature sampling[7]. However, fewer researchers study on how to measure features correctly, because character n-grams are compelling features for authorship attribution whenever researchers use term frequency(TF) and inverse document frequency(IDF) to measure them since these measurements are very tolerant to typos and non-standard usage[13].

The main contributions of this paper include:

1)We first explore the effects of different frequency measurements on authorship attribution for short texts. And we proposed a TNPF feature set for authorship attribution on short texts.

2) Adequate experiment evaluations on two real datasets Enron email and Amazon food review demonstrate the effectiveness of our approach. For example, the proposed approach could lead to 7.76% improvement of precision on Enron email compared to TF-IDF.

In the next section, we review related work. In Section 3, we present our TNPF strategy to build a feature set for classification. In Section 4, we show that TNPF feature set gets better performance and discuss results. We conclude in Section 5.

2 Related Work

Due to a large number of candidate authors, similarity-based approaches perform better on authorship attribution for short texts. [5] is the simplest similarity-based method which only calculates the Jaccard similarity between a given text and the profile texts of authors to find the most similar author. [7][8] try to solve the problem of a vast number of candidates in authorship attribution. [12] combine some factors such as 'emotion,' interjections, punctuation, abbreviations and other low-level features and want to find how they affect the authorship attribution. [11] builds a new kind of feature named k-signature for authors and proposes a flexible pattern. [2] first consider un-content factor for authorship attribution and proves the time factor can improve the classification result.[3][4] focus on the authorship attribution for emails.

3 The proposed approach

3.1 Term frequency and Inverse document frequency

Authorship attribution for short texts is a kind of text classification. Current methods usually do not explain precisely which measurement of frequency they used for features. However, it also means that they often use standard measurements of text classification like term frequency and inverse document frequency for features. The former is based on the statistic. It is common sense that frequent feature should play a significant role for authorship attribution. The latter is often used with term frequency, which is inspired by information retrieval tasks(named TF-IDF). The TF-IDF of one feature in an author is defined as follows:

$$tf_{i,j} - idf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{N} n_{k,j}} \log \frac{|A|}{1 + \{j : t_i \in a_j\}}$$

where $n_{i,j}$ is the statistical frequency of the i_{th} feature in author j, $|A|$ is the number of authors, $j : t_i \in a_j$ is the number of authors which contain the i_{th} feature. After this process, the feature which is frequent in one author but not

common in all authors will be more significant for classification. Besides, they must get the statistical frequency first, before they build a TF-IDF feature set.

3.2 TNPF and 1F

The authors of short texts usually have a smaller feature set than long texts due to its length. [7] think that we do not know which features are useful for authorship attribution on short texts, and which are not. And they randomly choose a certain number of features from the full feature set for calculating similarity and repeat this process many times to find the most similar author.

We are inspired by this opinion, and our paper tries to figure out the question whether the statistical frequency could reflect the importance of features for authorship attribution. To answer this question, we propose two hypotheses. The first one is that the frequent features are still important, but their statistical frequency is not proper for authorship attribution to calculate similarity. We rank the feature set with statistical frequency and select top-N feature as the final feature set to make the size of feature set equally for each author and reduce the scale of the feature set. Then we use Position-Frequency to replace statistical frequency. The Position-Frequency of the i_{th} feature is defined as follows:

$$NP_i = N + 1 - i$$

where NP_i is the frequency of the i_{th} feature, N is the value of top-N. We name it as Top-N Position-Frequency(TNPF).

The second one is that all features play same roles in authorship attribution due to the limited length of short texts. We use number 1 as the frequency of all features and name it as 1-Frequency(1F).

Dataset	Enron Email	Amazon Food Review
Users	10	29
Avg.texts	344	192
Avg.length(characters)±stedv	129±55	263±27
Avg.character 4-grams±stedv	18901±5593	11901±4280

Table 1. Detail Properties of the Datasets

4 Experiments Evaluation

4.1 Datasets and Baselines

Experiments are performed on two real datasets: Enron email[6] and Amazon food review[10]. The first one is an email dataset that is usually used for authorship attribution on short texts[2]. The second one consists of 568454 food reviews and 256059 users from Amazon collected by Stanford. We do experiments on the

second one because we want to do a study on the performance when authors all talk about one topic.

We randomly selected about 300 emails from each of 10 prolific authors in Enron email dataset and about 200 reviews from each of 29 prolific authors in Amazon food review due to runtime constraints. As pre-processing, we deleted the forward content of each email, for it does not belong to the textual features of authors. We model texts and authors with character 4-grams because character 4-grams haven been proved to be the most efficient features in authorship attribution [9].After pre-processing, some statistics about the datasets are provided in Table 1.

A quick analysis of the data shows the different nature of the two datasets: although each author from two datasets has similar sum of characters,the authors of Amazon food review still have a smaller feature set than Enron email. Besides, the authors of Amazon food review have a more stable size of feature set among different authors than Enron email. It may suggest when people talk about one topic, they will prefer to choose some fixed words to express their opinions.

Enron Email	TF	TNPF	1F	**Amazon Food Review**	TF	TNPF	1F
Without IDF							
Total	0.5962	–	**0.6078**	Total	**0.6871**	–	0.6320
Top-1000	0.5920	**0.6567**	0.6327	Top-1000	0.6453	0.6913	**0.7062**
Top-5000	0.5956	**0.6924**	0.6600	Top-5000	0.6823	0.8052	**0.8101**
Top-15000	0.5929	**0.6975**	0.6723	Top-9000	0.6850	**0.8290**	0.7956
With IDF							
Total	**0.6882**	–	0.6834	Total	0.7268	–	**0.7710**
Top-1000	**0.6999**	0.6978	0.6825	Top-1000	0.5964	0.6092	**0.6372**
Top-5000	0.7098	**0.7494**	0.7149	Top-5000	0.6894	0.7235	**0.7561**
Top-15000	0.6900	**0.7649**	0.7290	top-9000	0.7049	0.7572	**0.7651**

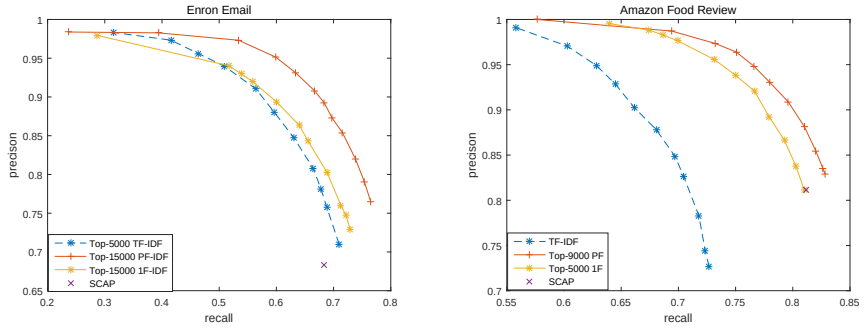
Table 2. The prediction accuracy of featuring sampling method using different frequency measurements on Enron email and Amazon food review. The best feature set is highlighted in bold. The best overall configuration is highlighted in bold and underlined.

4.2 Experiments

*Systems.*Since we are only concerned about similarity-based methods, our first method is the feature sampling method[7]. This method models texts and authors using unigram language model consisting of character 4-grams. Then they select k1 features from the profile of each author and calculate cosine similarity of the given texts with each author. After repeating k2 times, the given text will be determined as the most similar author, if the similarity is more than a predefined threshold otherwise no one is assigned to the text. We used Term Frequency(TF), Top-N Position-Frequency(TNPF) and 1-Frequency(1F) to measure character 4-grams for calculating cosine similarity on two datasets. Besides, we applied IDF

on three kinds of measurement for comparison. To figure out the effect of the scale of the feature set, we also selected Top-N features with TF feature set and 1F feature set.

The second method is the SCAP method, which is the simplest similarity-based method. It uses character 4-grams to deal with datasets and Jaccard index to calculate the similarity between the given text and the profile texts of authors. Then the given text is assigned to the most similar author. We use it only for comparison because this method does not use frequency for calculating similarity. Besides, 8-fold cross-validation is carried out on two datasets.



(a) (b)

Fig. 1. Precison-Recall curves of TF,TNPF,IF feature sampling method on Enron email and Amazon Food Review datasets. SCAP is another method for comparison.

Experimental results. Results according to feature sampling method of TF, TNPF, 1F with and without IDF on two datasets are reported in Table 2. Due to the different scale of feature sets in two datasets, we use different values of N for selecting top-N features. We set threshold=0, so all test texts will be assigned to an author. The best performance is achieved with k1=0.4 and k2=100. Result differences among TNPF, TF and 1F were found to be statistically significant. As the results show, TNPF outperforms standard TF and 1F except when N=1000 with IDF. On Amazon food review, we find that 1F perform better than TNPF and TF. However, the best performance is still top-9000 Position-Frequency without IDF. Besides, IDF will lead to a negative effect on authorship attribution with TNPF and 1F but they still outperform TF-IDF. There is another conclusion that we can get from two datasets: the size of the feature set has a bigger effect on the performance of TNPF and 1F than classical TF.

To demonstrate that the performance of TNPF is stable, we draw the precision-recall curves of the best performance for each measurement on Enron email and Amazon food review by setting different values of the threshold. We use SCAP as another method for comparison. As result shows, the precision of TNPF is higher than other two measurements at different recall points on two datasets.

5 Conclusions and Future Work

To measure features properly for authorship attribution on short texts, we proposed TNPF, a new position-based frequency measurement for calculating similarity. We offered 1F for comparison. We showed an experimental evaluation that TNPF outperforms classical statistical frequency and 1F. It also convinces our first hypothesis: the frequent features are important, but their statistical frequency could not reflect how important the features are. We think the reason may be that short texts only have a few words which are not adequate for statistics. Besides, we can also conclude that it will lead to better performance for authorship attribution on short texts when people talk about one topic.

In the future, we want to find a new similarity algorithm to find the most similar author, because the current algorithms still treat the features without distinction except frequency.

References

1. Abbasi, A., Chen, H.: Writeprints:a stylometric approach to identity-level identification and similarity detection in cyberspace. Acm Transactions on Information Systems 26(2), 1–29 (2008)
2. Azarbonyad, H., Dehghani, M., Marx, M., Kamps, J.: Time-aware authorship attribution for short text streams. In: International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 727–730 (2015)
3. Brocardo, M.L., Traore, I., Saad, S., Woungang, I.: Authorship verification for short messages using stylometry. In: International Conference on Computer, Information and Telecommunication Systems. pp. 1–6 (2014)
4. Corney, M.W., Anderson, A.M., Mohay, G.M., Vel, O.D.: Identifying the authors of suspect email. Communications of the Acm (2001)
5. Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E., Howald, B.S.: Identifying authorship by byte-level n-grams: The source code author profile (scap) method. International Journal of Digital Evidence 6(1) (2007)
6. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. Lecture Notes in Computer Science 3201, 217–226 (2004)
7. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45(1), 83–94 (2011)
8. Koppel, M., Schler, J., Argamon, S., Messeri, E.: Authorship attribution with thousands of candidate authors. In: International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 659–660 (2006)
9. Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Cybercrime and Trustworthy Computing Workshop. pp. 1–8 (2010)
10. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (Jun 2014)
11. Schwartz, R., Tsur, O., Rappoport, A., Koppel, M.: Authorship attribution of micro-messages. EMLNP (2013)
12. Silva, R.S., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.: Automatic Authorship Analysis of Micro-Blogging Messages. Springer Berlin Heidelberg (2011)
13. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the Association for Information Science and Technology 60(3), 538–556 (2009)