Authorship Verification based on Lucene architecture

Zhihao Liao[1], Yong Han[1*], Leilei Kong[1], Zhuopeng Hong[1], Zijian Li[1], Guiyuan Liang[1], Zhenwei Mo[1], Zhixian Li[1], Zhongyuan Han[1]

[1] *Foshan University, Foshan, China*

Abstract

Authorship verification is the task of deciding whether two texts have been written by the same author based on comparing the texts' writing styles. We regard this task as a retrieval problem and a model based on information retrieval is proposed for verifying the authorship. The vector space model is used to estimate the simiarity between documents. To consider the different features of documents, we build four kinds of indexes for each document. Then a weighted score is computed to decide whether two documents come from the same author. Using this simple-minded approach, we get achieved 0.3032 on the overall score.

Keywords

Lucene, indexing, Retrieval

1. Introduction

This paper proposes an implementation method of author authentication shared task in PAN 2021. The goal of the task was to create a way to predict whether two given documents were written by the same person [2]. In the task, we used vector space modal to build indexes focused on different views. The vector space model is implemented by Lucene. In addition, we consider the different characteristics of the document, such as the influence of stoppage words on the discrimination, the influence of capitalization on the discrimination, the influence of nouns on the discrimination, and the influence of adjectives on the discrimination. Different characteristics are used to judge the similarity of the text. Then we retrieve the documents in the test set, and we get a ranking list of scores for each document. Because each document is indexed from four views, there are four ranking lists of scores for each document. We weighted the scores for each document to get a new ranking list of scores. Rank the new list of scores from highest to lowest. When we determine whether two papers were written by the same person, we compare whether the first place in the weighted score ranking list of the two papers is the same. If they are the same, they are written by the same person, otherwise they are not written by the same person.

2. Datasets

The data set provided for this task is from the English documentation of fanfiction.net. Each line in the training set gives two passages, which may be written by the same author or by different authors. In the Truth file, each line records whether the two documents were written by the same author. In this paper, I only used the Small training set, which has 52601 rows of data and a total of 52655 authors. In the test set provided by the task, the authors included in the training set did not appear, which made the task more difficult. The test set is given by the authorship verification shared task at PAN21.

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: liaozhihaoqwq@outlook.com (A. 1); kongleilei@fosu.edu.cn (A. 2)(*corresponding author)
ORCID: 0000-0003-4864-7156 (A. 1); 0000-0002-4636-3507(A. 2)

3. Method

This document describes the model we built to verify authorship. In the pre-processing, we do four kinds of pre-processing for each document, which are converted to lowercase letters, deleted stop words, retained nouns, and retained adjectives. Then we set up four indexes for the processed documents. We retrieve documents for the test set. A ranking list of scores retrieved by weighted summation. When determining whether two documents were written by the same person, if their highest score in the weighted ranking list is the same, then they were written by the same author; otherwise they are not.

When comparing whether document1 and document2 were written by the same author, we retrieved document1 and document2 respectively. For document1 and document2, we can get four score ranking lists respectively. We weight each of the four score ranking lists and then add them together to get a final score ranking list. Finally, we compare the final score ranking list for document1 and document2. If the author at the top of the final ranking list for document1 and document2 is the same, we assume that document1 and document2 were written by the same person; otherwise, they were not by the same person.

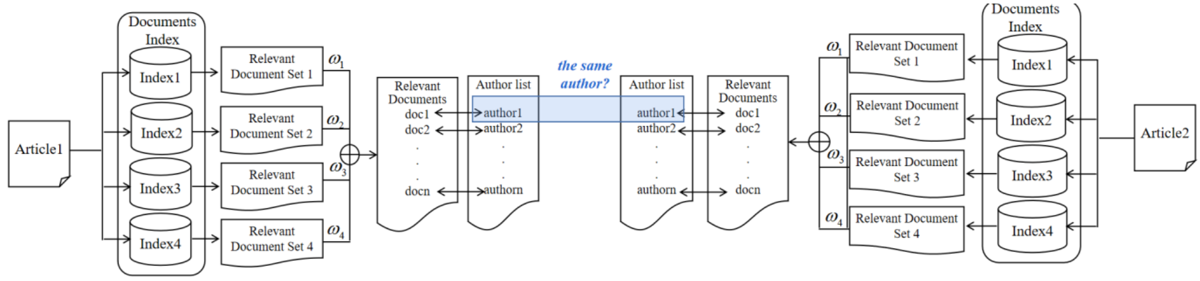

Figure 1: Retrieval result judgment diagram

For all the documents in the training set, we know which author wrote the document, so we first classified the documents in the training set according to their authors before establishing the index. For example, if text1 was written by author1, we would write text1 to file 1. If text2 is also written by author1, text2 will also be written to file 1. If text3 is written by author2, text3 will be written to file 3.

After we have classified the document, we can build the index. We first use the Lucene framework to index documents converted to lowercase letters [1]. Similarly, we index documents that remove stops, documents that retain nouns, and documents that retain adjectives.

After retrieving all the documents in the test set with each of the four indexes, we can get four score ranking lists for each document. We weighted the score ranking list of the four groups of authors to get the final score ranking list. In the final score ranking list, if the highest score for document1 and document2 corresponds to author1, then we can assume that both document1 and document2 were written by author1 and that the value we write in the result file is 1. If not written by the same author, the result file has a value of 0.

The final scoring formula can be expressed as:

$$\begin{aligned} final\_score(author) \\ = score\_lower(author)*\omega_1 + score\_nostopword(author)*\omega_2 \\ + score\_norn(author)*\omega_3 score\_adj(author)*\omega_4 \end{aligned} \tag{1}$$

The final_score(author) in the formula represents the probability score that an document was written by an author. Score_lower(author) represents the probability score that the lowercase model considers to be written by the author after an document is converted to lowercase. Score_nostopword(author) represents the probability score that the delete stop model thinks was written by the author after the stopword is deleted from an document. Score_norn(author) represents the probability score that the retention noun model considers to be written by the author after an document has retained a noun. Score_adj(author) represents the probability score that the retained adjective model considers to be written by the author after an document has retained an adjective. ω_1 , ω_2 , ω_3 and ω_4 are the weights set by the four models respectively.

4. Results

We deployed these program into the TIRA evaluation system provided by the PAN 2020 organizer, where the models were evaluated with unpublished data [6]. The system will compare and rank based on five supplementary indicators, namely AUC, F1-score, C@1, F_0.5U and BRIER.

Here are the results when we use the training set.

Table 1

Results of training set experiments

Method	F1	AUC	c@1	f_05_u	overall
baseline	0.785	0.808	0.743	0.71	0.762
M1-train	0.884	0.878	0.943	0.869	0.894
M2-train	0.885	0.897	0.891	0.951	0.906

M1-train represents $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 1$, $\omega_4 = 1$. M2-train represents $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 1.6$, $\omega_4 = 1.6$.

Baseline is a simple method based on text compression that given a pair of texts calculates the cross-entropy of text2 using the Prediction by Partial Matching model of text1 and vice-versa. Then, the mean and absolute difference of the two cross-entropies are used by a logistic regression model to estimate a verification score in [0,1].

Here are the results released by PAN on the test set [2].

Table 2

Test set experimental results

Method	F1	AUC	c@1	f_05_u	Brier	overall
M2-test-small	0.0067	0.4962	0.4962	0.0161	0.4962	overall

M2-test-small represents $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 1.6$, $\omega_4 = 1.6$.

5. Conclusion

Authorship verification is the task of deciding whether two texts have been written by the same author based on comparing the texts' writing styles. In order to complete the task of author authentication, this paper proposes to use vector space modal to establish indexes to classify author attribution. The vector space model is implemented by Lucene. To consider the different features of documents, we build four kinds of indexes for each document. This method has been experimented on the test set provided by the Authors Verification Shared Task of PAN21, and the result is F1 = 0.0067, auc = 0.4962, c@1 = 0.4962, f_05_u = 0.0161, Brier = 0.4962, overall = 0.3032.

In addition, we can see from the result, we just choose the characteristics of the nouns and adjectives as a writer is not very accurate, the follow-up work, we should use a more effective method to extract the characteristics of each document, compare the characteristics of the two documents are similar, so as to determine whether two documents written by the same authors.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61806075 and No.61772177), and the Social Science Foundation of Guangdong Province (No. GD20CTS02).

7. References

[1] GUAN Jian-he, GAN Jian-feng.: Design and implementation of web search engine based on Lucene. Computer Engineering and Design (2007)

[2] Kestemont, M., Markov, I., Stamatatos, E., Manjavacas, E., Bevendorff, J., Potthast, M. and Stein, B.: Overview of the Authorship Verification Task at PAN 2021. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR-WS.org (2021)