Similarity Ranking as an Attribute for Machine Learning Approach to Authorship Identification

Jan Rygl, Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
{xrygl,hales}@fi.muni.cz

Abstract

In the authorship identification task, examples of short writings of N authors and an anonymous document written by one of these N authors are given. The task is to determine the authorship of the anonymous text. Practically all approaches solved this problem with machine learning methods. The input attributes for the machine learning process are usually formed by stylistic or grammatical properties of individual documents or a defined similarity between a document and an author. In this paper, we present the results of an experiment to extend the machine learning attributes by ranking the similarity between a document and an author: we transform the similarity between an unknown document and one of the N authors to the order in which the author is the most similar to the document in the set of N authors. The comparison of similarity probability and similarity ranking was made using the Support Vector Machines algorithm. The results show that machine learning methods perform slightly better with attributes based on the ranking of similarity than with previously used similarity between an author and a document.

Keywords: authorship identification, machine learning, similarity ranking

1. Introduction

One of the current public safety challenges lies in intelligent monitoring of online media for extremist group communications. Since the authors of such contributions frequently hide themselves behind pseudonyms, there is a need of revealing the identity of anonymous writers (Abbasi and Chen, 2005; Chen et al., 2002). Our study addresses the anonymity problem by extending techniques solving the most typical scenario given as:

> Let us have an document d written by an unknown author and N groups of documents, where each group is written by one selected author (i.e. we have also N different authors). The goal is to assign the document d to the group in which the documents have the same author as d (Mosteller and Wallace, 1964).

2. Authorship Identification

Essentially, (Mosteller and Wallace, 1964) initiated authorship attribution studies. Since then much research was done on this topic taking the advantage of new results in areas such as machine learning, information retrieval, or natural language processing. One of the most important approaches to authorship identification lies in similarity-based models (proposed by (Burrows, 2002)).

Currently prevailing techniques use machine learning that works in two steps. First, documents with known authorship are divided into groups representing *authors' documents* and a test set acting as *unknown documents*. Positive and negative examples are created measuring similarities between *authors'* and *unknown documents*. Then a machine learning model is built. The second step consists of counting similarities between each possible author and the document with unknown authorship. For each pair of a document and an author, the similarity is calculated and converted to the overall probability of the same authorship using the model learned in the first step. Finally, the author with the highest *probability of the authorship* is selected as the author of the anonymous document.

Authors' *characteristic functions* are used to quantify the similarity between an author and a document. Each *characteristic* associates the document and author's group with their resemblance according to one of several criteria. All used characteristics are described in Table 1, they are sorted by their accuracy which was measured in (Rygl, 2011). The accuracy is measured using 250 training documents and 250 test documents written by four authors of Czech blogs. The values are higher than the values obtained in the result section because:

1. the measurement was made using only four authors (problem baseline is $\frac{1}{\textit{number of authors}}$)
2. all documents are of the same type (blogs) and from one domain (presented final results are measured using blogs, forum posts and blog comments from different Internet websites)
3. the number of training documents is significantly bigger (usually we are given only several authors' documents and our task is to find other documents of the same author)

The resulting resemblances are expressed as rational numbers. To sum up, the list of outputs of characteristic functions serves as input attributes for the machine learning process.

Authors are labeled as A and they are represented by their sets of documents. Characteristic functions are labeled as C and documents with unknown authorship as d.

method name	description	correctly classified
statistics of the morphological tags	*Morphological tags are extracted using Ajka (Sedláček, 2005).*	42 %
frequency of word classes	*Word classes are extracted using Ajka (Sedláček, 2005).*	33 %
statistics on punctuation	*Frequencies of punctuation symbols are compared, see (Chaski, 2005). Czech syntax analyser: (Jakubíček and Horák, 2010).*	26 %
frequency of bigrams	*Frequencies of bigrams are compared.*	26 %
statistics on the length of sentences	*Average length of sentences (Moritz, 1904)*	19 %
statistics on the length of words	*Frequencies of word lengths are compared.*	19 %
author's narrative style	*Gender and other statistics are extracted from the text.*	19 %
delta score	*Based on corpus word frequencies see (Stein and Argamon, 2006).*	17 %
statistics on typography	*Frequencies of standard typographical errors.*	17 %
word richness	*Ratio of unique words in the text (Holmes, 1985, p. 334).*	16 %
frequency of stop words	*Frequencies of the most common short function words.*	15 %
statistics on the count of sentences	*Frequencies of sentence lengths*	12 %

Table 1: Author's characteristics used in the presented system.

For example let us have authors A_{Adam} and A_{John}, an anonymous document d and characteristics $C_{delta\ score}$ and $C_{word\ richness}$. To compare the anonymous document to authors, we compute the similarity between authors and documents. Similarity between an author and a document is defined as an n-tuple (n is the number of characteristics) of similarity scores according to the characteristics:

$$\begin{aligned} Sim(d, A_{Adam}) &= (C_{delta\ score}(d, A_{Adam}), \\ &\quad C_{word\ richness}(d, A_{Adam})) \\ Sim(d, A_{John}) &= (C_{delta\ score}(d, A_{John}), \\ &\quad C_{word\ richness}(d, A_{John})) \end{aligned}$$

The machine learning classificator transforms tuples to probabilities of the same authorship and the author with the highest probability is selected. Despite the fact that characteristic functions define the similarity between authors and documents, the machine learning step is necessary to achieve reliable and precise results. Some characteristics fail on short documents[1] and some others need significantly more training documents. The combination of all characteristics allow this method to be used universally.

3. Authors Positions as Similarity Factor

We claim that the authorship detection is improved if we replace *similarities* between an author and a document (according to the characteristics) with the author's positions in *rankings* (generated from these similarities).

For example if we are given a set of three authors A_1, A_2, A_3, an anonymous document d and the characteristic C, instead of scores:

$$\begin{aligned} C(d, A_1) &= 0.5, \\ C(d, A_2) &= 0.7, \\ C(d, A_3) &= 0.2, \end{aligned}$$

the ranking function R:

$$\begin{aligned} R(d, A_1) &= 2, \\ R(d, A_2) &= 1, \\ R(d, A_3) &= 3 \end{aligned}$$

is used as an input for the machine learning.

If we consider the problem of authorship identification as a "competition" among potential authors, we can use a sport analogy: If athletes compete in the same weather, the same health conditions and the track is always the same, we recognize the best athlete by his or her score (time, points, etc.). But the best athlete is not necessarily the holder of the best score. What matters is the position of athletes in rankings. This compensates for unequal (real) conditions at each competition. To explain why we need to consider unequal conditions, we will have a look at two example sets of documents.

In the first set, let us have documents talking about one topic and written by authors A_1, A_2, A_3. We also have an anonymous document d_{set_1} talking about the same topic. Due to the shared topic, the documents contain many similar words. That affects the calculated characteristics (the scores are high in the normalized interval $\langle 0, 1 \rangle$), e.g.

$$\begin{aligned} C(d_{set_1}, A_1) &= 0.8, \\ C(d_{set_1}, A_2) &= 0.7, \\ C(d_{set_1}, A_3) &= 0.9. \end{aligned}$$

Now let us take another set that contains documents of different lengths and topics. The authors of the documents in this second set are the same as in the first set. Most of the characteristics depend on the statistics of various phenomena in the text. If documents are of various length and topic, then the respective phenomena must occur in varying degrees, thus reducing the overall scores to e.g.

$$\begin{aligned} C(d_{set_2}, A_1) &= 0.4, \\ C(d_{set_2}, A_2) &= 0.3, \\ C(d_{set_2}, A_3) &= 0.5. \end{aligned}$$

[1] We mean that they do not provide the requested demonstrative evidence, not that they would somehow fail to be computed.

	training the SVM model the training data	testing the SVM model authors' groups	unknown documents
documents	100 ($d_1, \ldots, d_{100}$)	100 ($d_1, \ldots, d_{100}$)	100 ($d_{101}, \ldots, d_{200}$)
authors	16 (A,...,P)	16 (A,...,P)	16 (A,...,P)
doc. per author	10+	5+	5+

method	accuracy	baseline	relative improvement
Similarity score	8%	6.25%	+28%
Position in the ranking	11%	6.25%	+**76**%
Combination of both	9%	6.25%	+43%

Table 2: In this table, a new model is built for each task. Authors' documents are used to create a machine learning model – the model is tuned to the examined data, but it is impossible to classify unknown documents if we are given insufficient number of authors' documents.

	training the SVM model the training data	testing the SVM model authors' groups	unknown documents
documents	100 ($d_1, \ldots, d_{100}$)	100 ($d_{101}, \ldots, d_{200}$)	100 ($d_{201}, \ldots, d_{300}$)
authors	16 (A,...,P)	10 (Q,...,Y)	10 (Q,...,Y)
doc. per author	10+	10+	10+

method	accuracy	baseline	relative improvement
Similarity score	12%	10.0%	+20%
Position in the ranking	17%	10.0%	+**70**%
Combination of both	14%	10.0%	+40%

Table 3: This table displays the situation when the trained model is independent of the task. The machine learning process is trained only once and the task is independent of the number of authors' documents

The problem is that values 0.9 and 0.5 indicate the same authorship and the values 0.7, 0.8, 0.4 and 0.3 are used for the different authorship. The machine learning can deal with this problem, but at the cost of reduced accuracy.

To further optimize the machine learning process, we have changed the ranking function R to the inverse function $S = \frac{1}{R}$. Therefore, the ordinal values $1, 2, 3, \ldots, N$ are transformed to the values of $1, 1/2, 1/3, \ldots, 1/N$. The main advantage of this inverse function S is that there is not such a big difference between problems with different numbers of authors. The best authors are evaluated equally and the worst authors' values differ absolutely by a small amount when compared to the difference between the first positions (i.e. $\frac{1}{\textit{number going to}N}$). This means that the training examples of a problem containing 5 authors are applicable to a problem containing 20 authors.

The suggested function S is consistent across different situations and returns values in the interval $\langle 0, 1 \rangle$, which is recommended for the implementation of Support Vector Machines algorithm[2] (Hsu et al., 2010) that we use in the authorship identification task.

4. Experimental results

To evaluate the ranking attribute approach, documents from the corpus of Czech texts *CzAu*[3] were used. The corpus consists of discussion posts containing at least three sentences and blogs freely available on the Internet. Texts are preprocessed by automatic morphological tools in the sense that they are tokenized, segmented and morphologically annotated (Sedláček, 2005).

A Library for Support Vector Machines (Chang and Lin, 2001) was selected for the machine learning task. The model builder received 12 attributes as input, which were computed using 12 author characteristic.

Two tests were performed to compare the similarity score (the characteristic function $C(\textit{author}, \textit{unknown document})$) to the ordered position in the ranking of characteristic score, the function $S = 1/(\textit{position of}\, C)$. The first test solved an easier task when there is enough data available to create a machine learning model. The task and results are described in Table 2. The second test simulates a situation when the lack of data for the tested authors forces us to use a model trained on documents by a different group of authors. The results are presented in Table 3.

Although the new function S does not bring great improvement in absolute terms, the relative success rate is increased significantly. The new ranking scores exceed similarity scores by 50 % when compared to the problem's baseline (results given by a naive random algorithm, in such case the baseline is counted as $\frac{1}{\textit{number-of-authors}}$). All scores increase when long documents are used.

To achieve further improvements, we combined the position in ranking with the similarity attributes. Since existing machine learning methods work with up to thousands of attributes, we could easily add new information for the decision making by doubling the number of attributes.

[2]Support Vector Machines algorithm is the most frequently used method to solve authorship identification problems (Koppel et al., 2009).

[3]The corpus is part of the project *Analysis of natural language on the Internet* and it can not be published.

However, the presented evaluation showed that the machine learning process provides better results with small amount of quality attributes when there are are only a few training documents available.

5. Conclusions and future work

We have shown that positions in rankings provide better results than similarities between documents and authors. The authorship identification problem is very difficult, therefore, we can consider each absolute accuracy increase to be a success if the relative increase compared to the problem baseline is high. Furthermore, it appears that for problems with a limited number of training data it is advisable to use only high quality attributes at the cost of less information being used to obtain better authorship classification.

In the future, experiments will be conducted on a larger scale. Also documents written in other languages will be used. The performance of attributes in other scenarios needs to be evaluated, e.g. different number of author groups and other types of documents. Most importantly, modifications of suggested attributes and various combinations of attributes will be tested to improve the accuracy.

6. Acknowledgements

This work has been partly supported by the Ministry of the Interior of CR within the project VF20102014003. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Ministry of the Interior of CR within the project VF20102014003.

7. References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

John Burrows. 2002. Delta': a measure of stylistic authorship 1. *Literary and Linguistic Computing*, 17(3):267–287.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. URL: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C. E. Chaski. 2005. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 4(1):1–13.

H. Chen, H. Atabakhsh, D. Zeng, et al. 2002. COPLINK: visualization and collaboration for law enforcement. In *Proceedings of the 2002 annual national conference on Digital government research*, dg.o '02, pages 1–7. Digital Government Society of North America.

D. I. Holmes. 1985. The Analysis of Literary Style–A Review. *Journal of the Royal Statistical Society*, 148(4):328–341.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2010. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan. URL: `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

Miloš Jakubíček and Aleš Horák. 2010. Punctuation Detection with Full Syntactic Parsing. *Research in Computing Science, Special issue: Natural Language Processing and its Applications*, 46:335–343.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60:9–26, January.

R. E. Moritz. 1904. On the significance of characteristics curves of composition. *The Popular Science*, 6:132–147.

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

J. Rygl. 2011. Determining Authorship of Anonymous Texts Based on Automatically Discovered Characteristic Features (Czech language). Master's thesis, Masaryk University.

Radek Sedláček. 2005. *Morphemic Analyser for Czech*. Ph.D. thesis, Faculty of Informatics Masaryk University.

Sterling Stein and Shlomo Argamon. 2006. A Mathematical Explanation of Burrows's Delta. Technical report, Linguistic Cognition Laboratory, Illinois Institute of Technology. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.8771&rep=rep1&type=pdf`.