Dynamic features selection in authorship identification problem

Anton Paramonov, Ilya Trukhanovich, Uladzimir Kuntsevich
Belarusian State University Informatics and Radioelectronics
Minsk, Republic of Belarus
ilya.trukhanovich@gmail.com, vkyntsevich@gmail.com

***Abstract*—The work is devoted to the problems of feature selection in the method of authorship identification in the context of authorial invariant defining. The problem of text author identification and existing approaches are described. In this article, models, methods, and experiment results for implementation of the dynamic method of defining features by genetic algorithm.**

***Keywords*—text author identification, texts classification, authorial invariant, genetic algorithm, support vector machine**

I. Introduction

The task of attribution of unknown text is an important information problem.

This is mainly due to the widespread use of messaging programs on the Internet, the increasing importance of e-mail for corporate correspondence, the popularity of forums and blogs. Without registering, users can send messages and specify their own information and registration is frequently simply symbolic. It is the same for e-mails and messengers. It means that the registration data cannot identify the contact person unambiguously, the address of the sender can be easily changed. Increasingly, the anonymity of internet messages is attracting cyberspace criminals [1].

In other areas such methods can also be applied. These techniques can be used in linguistic research to study authorship phenomena. The difference in this or that writer's style is of interest. Features that make his speech as individual or common characteristics of any characteristics easily visible. Authorship is unknown in a number of unassigned literary texts. It is evident that the existence of precise quantitative methods for the identification and evaluation of the author can resolve most controversial historical and literary criticism issues [2].

Education is another field of application. Students tend to do their own tasks less and prefer to spend less time and use prepared results. A more objective assessment approach will in this case be possible by using attribution methods [3].

The area uses machine learning, information retrieval and natural language processing approaches in a relatively interdisciplinary way. The authorial invariant is usually used as a "handwriting". This is a quantitative feature of literary texts or a parameter that uniquely characterizes one author's work or a small number of "closer authors" by their behavior and values in different groups of authors [4].

II. Problem Statement

The problem is formulated as follows when identifying the author of a text with a small number of alternatives. Assume that we have texts set $T = \{t_1, ..., t_n\}$ and authors set $A = \{a_1, ..., a_k\}$. We know authors for some subset $T_1 = \{t_1, ..., t_l\} \subseteq T$ so we have pairs like $\{a_i, t_j\}$ $(i = 1..k, j = 1..l)$ as training set L. The true authors of the remaining texts subset T_2 must be established.

In this context the problem of authorship can be seen as a problem in several classes. In this case, L is training set, A is a set of predefined classes, and T_2 is objects for classification. The aim is to create a classifier that solves this problem i.e. find a function that gives its true author an arbitrary text of set T.

A sequence of the following actions is part of the general technique for identifying the author of an unknown text:

- Selecting a text model in the form of feature sets.
- Select a group of characteristics for checking and forming an invariant of the author.
- The selection and the parameters for the classifier.
- Formation of an author's style model allowing two or more authors to be separated by a trained and invariant author.
- The authorship of the unknown text will be determined directly.
- The final decision by the classifier on the author of the text should be adopted if several groups of informative text features could be found.

You can use the word bag model and the n-gram one to represent texts in the information system. The word model bag is a collection of all words (or word attributes) which compose the text unordered. Text is defined as a sequence of n-element strings in the N-gram model [5].

The following sequence of measures is proposed to determine the differences in the authors' styles:

- Division into two groups of the existing set of texts. The first is for training the model classifier. Secondly, the identification author's accuracy is checked by using a trained model.
- Formation in accordance with the selected model of text representation in the form of a set of characteristics of a text vector characteristics from the invariant of the obtained author.
- Run attribute values into a single range through standardization and scaling operations.
- Correction by training the classifier in the normalized vectors of the features of the training text groups to ensure a high degrees of separation capability of the authors and verifying the accuracy of the trained classifier in the feature vectors of the test texts group.
- Change the list of the characteristics and/or property groups that constitute the group if it is not possible to achieve acceptable results by changing the parameters of the classification system.

III. Existing approaches

At the moment, a certain number of different solutions have been worked out.

K-means, support vector machine, neural network and other approaches were used as classifier basis.

Relevance of SVM as classifier basis was proved significantly by Thorsten Joachims [6]. Later it was tested by Rong Zheng and his colleagues [7].

Also neural network approach becomes more popular, especially in recent years, because of scientists from Stanford University [8] [9]. But it still requires significant improvements because of low accuracy in different cases and a lot of resources to select architecture and train.

The question of the set of features that make up the author's invariant is under discussion. Usually approaches are based on stylistics, syntactic, lexical and other features.

But nowadays a lot of approaches use low-level features like punctuation frequency, average length and so on and they works rather well with above 80% accuracy, It was demonstrated by scientists of Pace University and University of Sheffield [10] [11]. Besides we should admit that feature set for literary books is significantly determined. The fact is widely used in papers like work of University of Ottawa [12]. But for common cases it is still full of uncertainty.

IV. Supposed approach

Let's define key parameters of supposed approach. Text can be represented as feature set and success of classification depends on feature combinations quality.

Lexical, syntactic, structural, content-specific, style and other features can be used as author's invariant. Despite the use of completely different levels of abstraction, completely different ones can give acceptable results within the scope of the classification.

Using features from higher levels of hierarchy, the analyzes of the structure of the text are made more complicated and difficult to automate with each new level. Due to the level of noise in the analysed text, the language characteristics, etc, and others, for example, inaccuracies may arise at each stage, leading to serious errors at higher analysis levels. It was therefore decided in this study to focus on the characteristics of the levels of chars and words.

Along with all this, as mentioned above, at present the question of the conventional set of features that make up the author's invariant remains open. Due to the mentioned above reasons this is caused by a wide range of applied problems, and as a consequence by a wide variety of texts domains. Features of a certain level of abstraction can work well within the scope of differences in one type of text (style features during classifying literary books), but rather poorly when trying to classify analytical reviews in blogs.

Thus, it seems theoretically justified to test an approach that, within the scope of the task and the proposed texts, allows you to select the optimal situational features. In this work, it is proposed to use a genetic algorithm.

The algorithm consist of the next steps:

- Create initial set of supposed features.
- Enter the loop
- Add new feature to the set
- Run genetic algorithm and select appropriate features
- If selected features have appropriate accuracy exit the loop
- Use selected features

V. Experiment

A. Initial parameters

We can take arbitrary Reuters news feed articles pack for for research purposes. We will take 2, 5 and 10 authors from the dataset and 50 texts for each one. We will include articles of known authors in test set.

We can use Support Vector Machine (SVM) as classificator basement. The SVM is a supervised learning machine algorithm that can be used for both classification and regression challenges. But it is mostly used in problems with classification. The value of each feature is a value in a given coordination, so that each data element is drawn as a point in N-dimensional space in the SVM algorithm, where n is the number of features. Then we classify by finding the hyperplane that very well distinguishes classes.

Initial features can be selected based on those that have already proven themselves well (table I).

Table I
INITIAL FEATURES

Feature name	Feature description
DICTIONARY-M	M most frequent words from language dictionary
WORDS-M	M most frequent words from sample dictionary
UNIGRAM-M	M most frequent unigrams
BIGRAM-M	M most frequent bigrams
TRIGRAM-M	M most frequent trigrams
TETRAGRAM-M	M most frequent tetragrams
PENTAGRAM-M	M most frequent pentagrams
POS	Parts of speech frequency
AVERAGE-WORD	Average word length in characters
AVERAGE-SENTENCE	Average sentence length in words

Fig. 1 displays dependence of accuracy for 5 authors on dimension for some features (DICTIONARY (blue), WORDS (green), TRIGRAM (orange)).

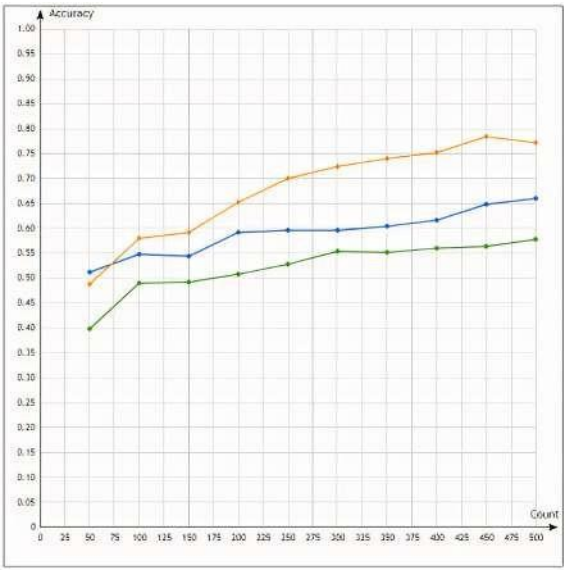

Figure 1. Dependence of accuracy for 5 authors (one feature)

Combining of features is associated with various effects. Let's take a look at this (Fig. 2).

Blue points are accuracy values for trigrams, gray ones are values for tetragrams. Orange points are values for combinations of trigrams and tetragrams. We can see significant increase of accuracy for some orange points but at the same time some ones show lower accuracy. That's why it is so important to choose feature combination carefully.

B. Results

According to experimental results we have 98% average accuracy for 2 authors, 84% accuracy for 5 authors and 65% accuracy for 10 authors. Best results uses DICTIONARY-M, WORDS-M and N-GRAM features (table II).

The results can be explained by the specificity of supposed texts.

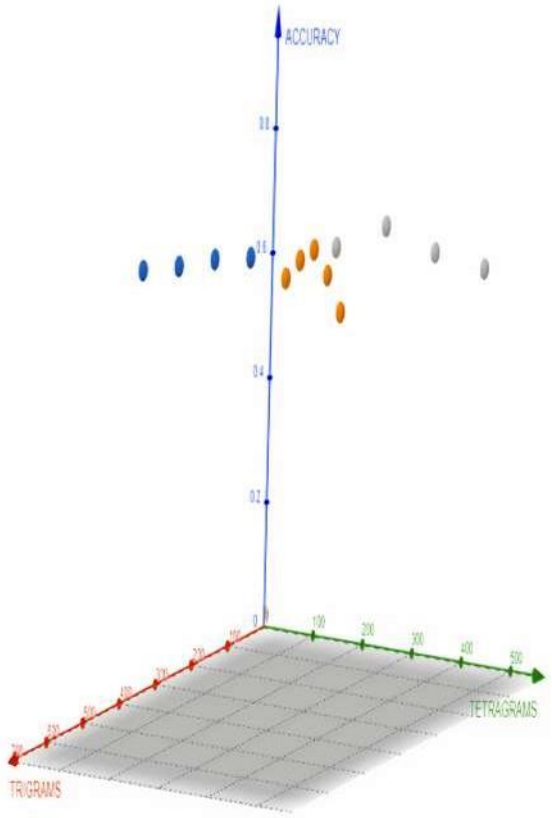

Figure 2. Dependence of accuracy for 5 authors (several features)

Table II
RESULTS

Number of authors	Average accuracy (test set)
2	98%
5	84%
10	65%

One of the most interesting features are expressive means but we didn't include them in initial feature set because they are not very relevant for news articles.

Words are the basis for authorial invariant so it is not surprising these features were selected. They underline text nature especially in our cases. At the same time N-GRAM feature is usually highly recommended and we can see that in our experiments they are relevant.

Other features are not so effective because, for example, parts of speech frequency does not shows personality for proposed news texts well because of their topics.

VI. CONCLUSION

Introduced approach has great potential because of its flexibility. Despite the fact that some aspects of text authorship are investigated a lot of patterns are not noticeable. Moreover, unexpected pattern combinations

can lead to efficient increasing of accuracy. After feature selecting we can investigate their influence after the fact.

And this approach still have growth points:

- Selecting classifier basis along with features
- Integrating as a part in hybrid classifier system
- Using several selecting rounds with different abstraction levels of features
- Analysing threshold values for determining unknown authors

References

[1] R. Zheng, Y. Qin, Z. Huang Authorship Analysis in Cybercrime Investigation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2003, vol. 2665, pp. 59-73.

[2] R. Griffin Anonymity and Authorship. *New Literary History*, 1999, vol. 30, no 4, pp. 877-895.

[3] Types of plagiarism. Available at: https://www.bradford.ac.uk/library/find-out-about/plagiarism/types-of-plagiarism/ (accessed 2021, Mar)

[4] Writer invariant definition. Available at: https://www.encyclo.co.uk/meaning-of-Writer_invariant (accessed 2021, Mar)

[5] A. Leonova, I. Leonova Opredelenie avtorstva tekstov na osnove podkhoda n-gramm [Determination of the n- gram-based text authorship] Available at: https://science-engineering.ru/ru/article/view?id=1205 (accessed 2021, Mar)

[6] T. Joachims Text Categorization With Support Vector Machines: Learning With Many Relevant Features, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 1998, vol. 1398, pp. 137-142.

[7] R. Zheng, J. Li, Z. Huang A framework for authorship analysis of online messages: Writing-style features and techniques, *Journal of the American Society for Information Science and Technology*, 2006, vol. 57, no 3, pp. 378-393.

[8] Deep Learning based Authorship Identification. Available at: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf (accessed 2021, Apr)

[9] Wallace: Author Detection via Recurrent Neural Networks. Available at: https://cs224d.stanford.edu/reports/YaoLeon.pdf (accessed 2021, Apr)

[10] Stylometry for E-mail Author Identification and Authentication. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.547.438&rep=rep1&type=pdf (accessed 2021, May)

[11] Topic or Style? Exploring the Most Useful Features for Authorship Attribution. Available at: https://www.aclweb.org/anthology/C18-1029/ (accessed 2021, May)

[12] Authorship Identification for Literary Book Recommendations. Available at: https://www.aclweb.org/anthology/C18-1033.pdf (accessed 2021, May)

Динамический выбор признаков в задаче идентификации автора

Парамонов А.И., Труханович И.А.,
Кунцевич В.С.

Работа посвящена проблемам выделения признаков в методе идентификации авторства в контексте определения авторского инварианта. Описана проблема идентификации автора текста и существующие подходы. В этой статье представлены модели, методы и результаты экспериментов по реализации динамического метода определения признаков с помощью генетического алгоритма.

Received 31.05.2021