TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
1-AC/2013

SZYMON ŁUKASIK*, MARCIN HARĘZA**, MARCIN KACZOR**

DOCUMENT CONTENT MINING FOR AUTHORS' IDENTIFICATION TASK

EKSPLORACJA TREŚCI DOKUMENTÓW W PROBLEMIE IDENTYFIKACJI AUTORÓW

Abstract

This paper deals with automatic authorship attribution through documents content analysis. This approach is based on selecting sets of suitable features relying on specific use of grammar, punctuation or vocabulary and in the next step – executing given classification algorithm. The contribution first overviews various text characteristics which can be employed for that purpose, then presents the results of experiments involving feature selection and examines classifier performance for author identification problem. The paper concludes with discussion and proposals for further research.

Keywords: author identification, feature selection, classification

Streszczenie

Przedmiotem niniejszego artykułu jest problem identyfikacji autora na podstawie analizy treści dokumentów. Podejście to opiera się na wyborze odpowiednich cech związanych ze specyficznym użyciem struktur gramatycznych, interpunkcji oraz słownika, a następnie – użycie wybranego algorytmu klasyfikacji. W artykule przedstawiono najpierw różne charakterystyki tekstu, które mogą być użyte w omawianym zagadnieniu, a następnie załączono wyniki eksperymentów obliczeniowych obejmujących wybór cech i badanie skuteczności klasyfikacji w problemie identyfikacji autorów. Artykuł podsumowano wnioskami oraz propozycjami dalszych prac w rozważanej tematyce badawczej.

Słowa kluczowe: identyfikacja autora, wybór cech, klasyfikacja

* Szymon Łukasik, Ph.D., Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology; Systems Research Institute, Polish Academy of Sciences.

** Marcin Haręza, Marcin Kaczor, Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology.

1. Introduction

Author identification is a task commonly performed in historical research, archeology and criminology. Historically it was predominantly considered in the context of handwritten text i.e. taking into account author's writing style. Nowadays, as most of documents are being stored in their electronic versions, it is impossible to complete the identification procedure employing solely graphical features of the text. Moreover, novel types of textual content like computer programming source code, pose new problems, by excluding the possibility of using graphical text representation in the framework of automatic authorship attribution.

The task of selecting author of a given text from a known list of authors can be perceived as a classification or pattern recognition problem. It is a commonly known issue in data mining [23], with a broad range of areas where it transpires, e.g. biometrics, medical diagnostics or intrusion detection systems [6, 14]. Classification is a task of assigning elements from so called testing set, denoted by matrix Y:

$$Y = \left[Y_1 \; Y_2 \; ... \; Y_n \right] \tag{1}$$

which n columns represent features of m_{test} objects belonging to this dataset, to one of the known C classes. Usually a set of representative elements for those classes is additionally given, in the form of a training dataset:

$$X = \left[X_1 \; X_2 \; ... \; X_n \right] \tag{2}$$

with m_{train} elements having class labels explicitly defined. The task for a classifier is to learn how to predict class assignment for testing dataset using knowledge acquired from training set. In case of the authors identification tasks, class label corresponds to author's identifier, training set to a set of documents with known authors and testing set to a group of documents which authorship needs to be identified. To successfully perform such tasks one should select capable classification algorithm and, even more importantly, define suitable document representation, in the form of n distinctive features.

This paper investigates a possibility of employing various non-graphical sets of features for author identification task. Among others we take into account presence of distinctive grammatical structures, quantitative analysis of parts of speech, and diversity of words or punctuation. It is usability of aforementioned features that is experimentally assessed for selected authors identification task.

There exists a vast amount of studies in the area of this contribution, however usually not--involving such a broad range of characteristics being examined at the same time. Previous work in this area involve using: lexical features (e.g. functional words), character features, including alphabetic or digit characters count, uppercase and lowercase characters, letter frequencies, etc. [5], syntactic features [11], semantic features [1] and application specific characteristics, like the use of greetings, signatures, etc. [24]. The problem of language specific issues is also widely studied [4]. Interesting applications of authorship attribution include microblogging posts author identification [15], gender recognition [3] or combining author classification with opinion mining [18]. Accomplished surveys of techniques and strategies commonly employed for authorship attribution tasks can be found in [9, 13, 22].

The paper is organized as follows. First, we give a detailed description of documents characteristics, useful for their representation in authorship attribution task. It is followed by experimental setup, employed to assess the efficiency of classification algorithm used with various documents features sets. Finally, the discussion and proposals for further research are given.

2. Document content representation

Document content will be described here by five groups of features corresponding to the following characteristics of a text, given in its computer-written representation:

a) document grammatical composition, represented by features' group G_R, composed of $R + 9$ characteristics $g_1, g_2, ..., g_{R+9}$ with R being a parameter related to the size of analyzed set of grammatical structures;
b) used punctuation marks, represented by features group P which includes 16 characteristics $p_1, p_2, ..., p_{16}$;
c) words length, given by features group $L : l_1, l_2, ..., l_4$;
d) document formatting defined by features group $F : f_1, f_2, ..., f_5$;
e) words usage described by features group $V : v_1, v_2, ..., v_9$.

The same set of features can be used to characterize numerous documents – belonging to both, training and testing datasets. Albeit, a document D description might be for example given by $D = \{G_R, V\}$ – it would mean that it is characterized only by indicators related to grammatical composition and used words statistics. The study of applicability of different set of features for author's identification task will be conducted in the next Section of the paper. We will now provide a description of features which were assigned to groups listed above.

2.1. Document grammatical composition

Grammar is a linguistic concept concerning both the shape of words and how words (and phrases) can be combined [2]. Grammar, in essence, consists of two components: morphology, i.e. study of how words are formed out of smaller units (morphemes) and the syntax, which is a system of rules specifying how lexical items ought to be composed together [21]. Of those two, when considering writing style analysis, the syntax is of more importance.

To analyze syntactic structure of sentences one should define a notation to conveniently represent content of a given text. Tree structure is commonly used for that purpose. Its branch nodes represent non-terminal syntactic symbols (with sentence as a root) and leaves are equivalent to lexical tokens of the sentence. To obtain structure in this form text needs to be parsed. This language-dependent task is one of the most important in natural language processing. Here, we assume that analyzed text corpus is parsed with popular Stanford parser [12].

To identify a presence of selected tree components, we established 256 most commonly used syntactic structures in English, extracted from the Wall Street Journal corpora contained by Penn Treebank project [19]. It can be found on the website [16]. On that basis, a feature set:

$$g_1, g_2, \ldots, g_R \tag{3}$$

was created. Each characteristic from this set indicates how many times top R structures from the ranking were used in the analyzed document. For example g_1 corresponds to the number of "preposition + noun phrase" structures found in the text. In general, for ranking size R one can assume any value between 1 and 256.

Set of features mentioned above corresponds to the presence of selected grammatical constructs in analyzed document. Next attribute g_{R+1} describes concentration of syntactic elements listed in the ranking. Let $N_R = \sum_{i=1}^{R} g_i$ denote occurrence of constructs from the selected set of R grammatical structures in the given text and let N_G to indicate overall number of structures pointed out by syntactic parser for the analyzed document. Then g_{R+1} can be written as:

$$g_{R+1} = \frac{N_R}{N_G} \tag{4}$$

Subsequent attributes $g_{R+2}, g_{R+3}, ..., g_{R+9}$ describe the occurrence of the individual parts of speech in the text. Here g_{R+2} corresponds to the incidence of nouns, g_{R+3} – pronouns, g_{R+4} – adjectives, g_{R+5} – verbs, g_{R+6} – adverbs, g_{R+7} – prepositions, g_{R+8} – conjunctions and g_{R+9} – interjections.

2.2. Document punctuation

Punctuation is a practice of inserting standardized signs to clarify the meaning and separate language structural units [20]. Among fourteen most commonly used punctuation marks in English, at parsing stage the following symbols are identified here: full stop, exclamation mark, question mark, comma, semicolon, colon, apostrophe, quotation mark, ellipsis, dash, hyphen, slash plus additionally multiple exclamation marks and multiple question marks. Use of those signs is indicated by features group P formed of 16 characteristics.

The first attribute p_1, denoted by:

$$p_1 = \frac{N_P}{N_C} \tag{5}$$

represents number of punctuation marks N_p listed above found in the text, divided by total number of characters N_c.

Second feature from this set p_2 describes variety of punctuation marks, by employing the following formula:

$$p_2 = N_P^* \tag{6}$$

where N_P^* symbolizes overall number of different punctuation signs found in the analyzed document.

Finally, features $p_3, p_4, ..., p_{16}$ refer to the number of times each of punctuation marks listed above was used in the text, scaled by total number of punctuation marks N_p.

2.3. Document word length statistics

Features group L aims at capturing authors' tendency to use short words, long words or in general – words of approximately similar length. Let us denote by S a set of all sentences in the text, and by W_s a set of all words forming a given sentence $s \in S$. Consequently, by size(w) where $w \in W_s$ we understand length of a selected word w while using card(S) and card(W_s) at the same time to represent cardinalities of both sets defined above.

First feature l_1 introduced here is referring to the average word length in the sentences forming the text, and can be written as:

$$l_1 = \frac{\sum_{s \in S} \frac{\sum_{w \in W_s} \text{size}(w)}{\text{card}(W_s)}}{\text{card}(S)} \tag{7}$$

Second feature l_2 describes average word length in the document and is represented by:

$$l_2 = \frac{\sum_{s \in S} \sum_{w \in W_s} \text{size}(w)}{\sum_{s \in S} \text{card}(S)} \tag{8}$$

Two other features from this group refer to the number of short words (consisting of less than 4 characters) and long words (made up of more than 6 characters). Both attributes are relative to the sum of words used in the text, and can be written as follows:

$$l_3 = \frac{\sum_{s \in S} \text{card}(W_s^{\text{short}})}{\sum_{s \in S} \text{card}(W_s)} \tag{9}$$

with $W_s^{\text{short}} = \{w : w \in W_s \wedge \text{size}(w) < 4\}$ and:

$$l_4 = \frac{\sum_{s \in S} \text{card}(W_s^{\text{long}})}{\sum_{s \in S} \text{card}(W_s)} \tag{10}$$

where $W_s^{\text{short}} = \{w : w \in W_s \wedge \text{size}(w) < 4\}$.

2.4. Document formatting

Formatted text, as opposed to plain text, introduces additional styling information. In general it can include colors, font, characters size or other special elements e.g. hyperlinks.

However, most of text repositories comprise only carriage returns or additional soft returns. The following set of features captures writers' individual preferences to use those formatting elements in their documents.

The first feature f_1 considered here, refers directly to the number of paragraphs, i.e. sections of the text with first line being indented. Next attribute f_2 captures average number of sentences included in one paragraph, which can be written as:

$$f_2 = \frac{\text{card}(S)}{N_{\text{par}}} \tag{11}$$

with N_{par} representing total number of paragraphs in the analyzed document.

Features f_3 and f_4 characterize average number of words and characters per paragraph. They can be written as follows:

$$f_3 = \frac{\sum_{s \in S} \text{card}(W_s)}{N_{\text{par}}} \tag{12}$$

and

$$f_4 = \frac{N_C}{N_{\text{par}}} \tag{13}$$

Finally, last feature from this group, f_5 describes writer's tendency to format the text with empty lines. It is represented by relative number of blank lines:

$$f_5 = \frac{N_L{}^E}{N_L} \tag{14}$$

with $N_L{}^E$ being total amount of empty lines in the text and N_L – all lines in the analyzed document.

2.5. Words use statistics

Authors usually differ in the sizes and structures of their vocabularies. Therefore the analysis of vocabulary richness and its concentration could be useful in the context of authorship attribution [7]. Feature set *V* being introduced here aims to quantitatively express those characteristics.

First feature under consideration v_1 refers directly to the number of distinct words used in the text N_v. Five features which follow, employ the concept of *hapax legomenon* (gr. said once) that is words that only occur once in the text [10] and *dis legomenon* – words appearing twice. Let us denote by N_{hl}, N_{dl} number of *hapaxes* found in the document. First five of aforementioned features are now defined as follows:

$$v_2 = \frac{N_{hl}}{\sum_{s \in S} \text{card}(W_s)} \tag{15}$$

$$v_3 = \frac{N_{hl}}{N_v} \tag{16}$$

$$v_4 = \frac{N_{dl}}{\sum_{s \in S} \text{card}(W_s)} \tag{17}$$

$$v_5 = \frac{N_{dl}}{N_v} \tag{18}$$

$$v_6 = \frac{100 \log_{10} \sum_{s \in S} \text{card}(W_s)}{1 - \frac{N_{hl}}{N_v}} \tag{19}$$

with the latter two known from the literature under the names of Sichel (v_5) and Honore (v_6) measures [3].

Defining additional text characteristics – N_{il} representing number of words occurring in the text precisely *i*-times, allows us to formulate three last features in this set, recognized as Yule (v_7), Simpson (v_8) and entropy (v_9) measures [3]:

$$v_7 = 10^4 \left[-\frac{1}{\sum_{s \in S} \text{card}(W_s)} + \sum_{i=1}^{N_v} N_{il} \left(\frac{i}{\sum_{s \in S} \text{card}(W_s)} \right)^2 \right] \tag{20}$$

$$v_8 = \sum_{i=1}^{N_v} N_{il} \frac{i}{\sum_{s \in S} \text{card}(W_s)} \frac{i-1}{\sum_{s \in S} \text{card}(W_s) - 1} \tag{21}$$

$$v_9 = \sum_{i=1}^{N_v} N_{il} \left(-\log_{10} \frac{i}{\sum_{s \in S} \text{card}(W_s)} \right) \frac{i}{\sum_{s \in S} \text{card}(W_s)} \tag{22}$$

3. Experimental studies

Experiments performed were designed to evaluate usefulness of features sets listed above for authorship attribution tasks and to study performance of classifiers employing carefully selected groups of attributes.

Experimental studies were conducted using Thomson Reuters Text Research Collection (known as TRC2 dataset). The dataset includes almost 2 million news reports collected between January 2008 and February 2009. We have chosen randomly documents authored

by 20 writers (around 100 contributions each). The first part consisting of 10 writers contributions was used for first part of experiments – involving feature set evaluation, and will be referred to as the training dataset. The rest of experimental data (labeled as the testing dataset) was employed in the second part of numerical studies, more thoroughly investigating the performance of selected classification algorithms.

As classifiers Naïve Bayes, feed-forward Neural Network with back-propagation learning, classic *k*-Nearest Neighbor (with $k = 3$) and Random Forests were chosen, with the latter two only being employed in the last part of this study. Optimal values of parameters for investigated techniques were determined through a set of pilot runs. For more details on those algorithms one could refer to [6].

3.1. Feature set selection

First we examined usability of features listed in the previous Section for authorship attribution. For that purpose Sequential Forward Search (SFS) or Sequential Backward Search (SBS) algorithm were executed, with *k*-nearest neighbor classifier being used for evaluation. Both techniques constitute similar supervised feature selection paradigms [17]. First algorithm starts with empty candidate feature subset and at each iteration adds the feature which maximizes classifier accuracy. SBS is the opposite strategy – it starts with the entire set of available features, and then iteratively removes one feature at the time, as long as aforementioned measure of accuracy improves, where the feature removed maximizes this improvement [8]. Here, the discriminative power of a given feature set was determined through cross-validation.

Feature selection algorithms were executed with $R = 256$ and five different schemes:

- Basic Feature Selection (BFS) – feature selection is performed on all groups of characteristics listed in Section 2.
- Two Phase Feature Selection (TPFS) – feature selection algorithm is executed on non-grammatical characteristics (all but G_R). Obtained reduced feature set is merged with G_R and feature selection is performed again.
- Parallel Feature Selection (PARFS) – feature selection is conducted in parallel on non-grammatical and grammatical characteristics, obtained reduced feature sets are merged afterwards.
- Grammatical Feature Selection (GFS) – feature selection is conducted only on characteristics from the grammatical ranking.
- Non-Grammatical Feature Selection (NGFS) – feature selection is executed only on characteristics not listed in grammatical ranking.

Table 1 illustrates a number of features obtained for different variants of feature selection. It is evidently observed that forward search principally results in more compact reduced set of characteristics, which is naturally caused by local search nature of this algorithm. Individual characteristics most frequently chosen in experiments involving different feature selection strategies, for both, features listed (top ten) and not listed (top ten) in the grammatical ranking are shown on Fig. 1. It can be observed that most useful features are found within grammatical ranking, especially between features g_{20} and g_{100}. Furthermore features describing incidence of pronouns (g_{R+3}), adverbs (g_{R+6}) and prepositions (g_{R+7}) appeared frequently in the reduced feature sets. Among non-grammatical characteristics ratio of punctuation marks and occurrence of exclamation mark, dash and hyphen were found particularly important; however in general selection ratio for those features seems to be significantly lower.

Table 1

Obtained features set sizes

Feature selection scheme	Sequential feature selection	
	Sequential Forward Search (SFS)	Sequential Backward Search (SBS)
BFS	20	222
TPFS	26	194
PARFS	37	240
GFS	24	233
NGFS	13	7

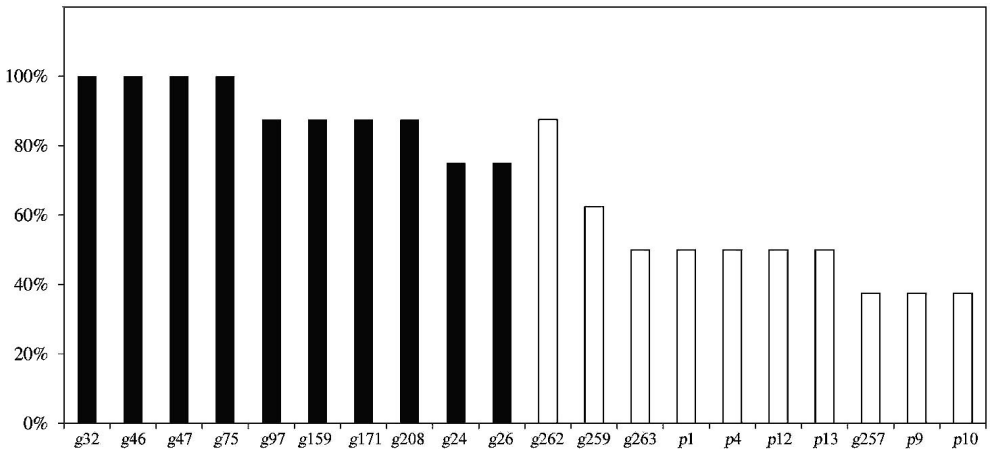

Fig. 1. Ten most frequently chosen characteristics for all variants of feature selection and features listed and not listed in the grammatical ranking

Rys. 1. Dziesięć najczęściej wybieranych charakterystyk dla wszystkich wariantów algorytmu wyboru cech i cech z i spoza rankingu gramatycznego

3.2. Performance analysis

Next series of experiments was devoted to studying performance of selected classifiers with different feature sets obtained through feature selection and varying size of grammatical ranking (where applicable). The trials were executed for training dataset first, to determine the best-performing classifier and the most appropriate feature set, with five runs involving cross-validation and average classification accuracy being reported here.

Tables 2–5 sum up obtained results for *k*-nearest neighbor, Naïve Bayes, feed-forward Neural Network and Random Forests classifiers.

Table 2

Classification accuracy [%] for training dataset and k-Nearest Neighbor classifier

Ranking Size & Sequential Selection Variant / Feature selection variant	*R* = 32		*R* = 64		*R* = 128		*R* = 256	
	SFS	SBS	**SFS**	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**
BFS	42.25	22.50	37.50	32.50	55.25	22.25	56.00	61.25
TPFS	33.00	19.50	46.75	39.75	43.50	26.75	53.75	39.50
PARFS	49.75	53.00	56.25	60.75	61.00	69.25	63.00	70.50
GFS	51.00	50.75	51.25	62.00	61.50	69.75	62.75	75.00
NGFS	38.25	20.25	32.75	22.50	25.00	20.25	38.50	24.25

Table 3

Classification accuracy [%] for training dataset and Naïve Bayes classifier

Ranking Size & Sequential Selection Variant / Feature selection variant	*R* = 32		*R* = 64		*R* = 128		*R* = 256	
	SFS	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**
BFS	37.50	24.00	27.25	22.50	40.50	18.75	39.75	58.75
TPFS	27.00	20.00	40.00	29.75	30.00	23.00	41.00	56.25
PARFS	42.25	42.25	45.75	59.00	48.75	62.25	50.50	55.75
GFS	38.00	44.00	41.75	57.75	43.75	62.25	45.50	56.25
NGFS	38.00	17.00	26.75	19.50	20.75	14.50	36.25	21.50

Table 4

Classification accuracy for training dataset and Neural Network classifier

Ranking Size & Sequential Selection Variant / Feature selection variant	*R* = 32		*R* = 64		*R* = 128		*R* = 256	
	SFS	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**
BFS	41.75	29.25	47.25	46.25	56.75	42.50	45.00	60.50
TPFS	27.75	26.25	48.25	41.25	50.00	33.25	48.75	60.50
PARFS	53.75	42.25	45.00	46.75	50.00	62.75	63.50	60.25
GFS	35.00	43.25	41.00	51.00	52.25	56.50	56.60	61.25
NGFS	38.50	17.00	26.25	11.75	24.25	17.25	35.00	22.25

Table 5

Classification accuracy for training dataset and Random Forests classifier

Ranking Size & Sequential Selection Variant / Feature selection variant	*R* = 32		*R* = 64		*R* = 128		*R* = 256	
	SFS	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**	**SFS**	**SBS**
BFS	66.00	33.50	57.25	51.50	69.25	58.75	68.00	82.00
TPFS	37.75	28.25	61.50	44.25	57.50	47.25	62.25	83.50
PARFS	62.75	64.75	60.25	69.00	69.25	80.50	76.00	79.25
GFS	61.75	63.75	53.75	70.75	71.00	80.50	68.00	80.75
NGFS	44.25	27.75	36.75	24.25	34.50	23.25	46.25	32.00

It can be seen that employing a broad range of grammatical features is crucial for obtaining high classification accuracy. Random Forests classifier was found to be the best performing one. Among various schemes of feature selection the most successful were Two Phase Feature Selection, Basic Feature Selection and Grammatical Feature Selection.

Finally selected best-performing classifiers based on Random Forests and *k*-Nearest Neighbor were more thoroughly evaluated for both datasets. We used set of characteristics with $R = 256$ and Sequential Backward Search along with Two Phase Feature Selection or Grammatical Feature Selection (for *k*-Nearest Neighbor). Results obtained for 10 runs and both training and testing datasets, with 5-fold cross-validation are shown on Fig. 2.

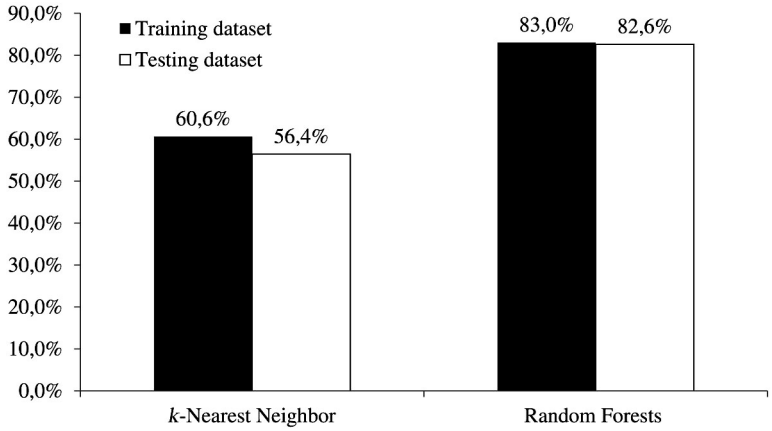

Fig. 2. Classification accuracy for training and testing datasets

Rys. 2. Trafność klasyfikacji dla zbiorów uczącego i testującego

4. Conclusion

This contribution examined the possibility of employing various characteristics of documents in computer-written form for authorship attribution. The suitability of several features was considered for parsed instances of news reports, taking into account a possibility of using few discrimination algorithms for conclusive classification. It was established here, that for ensuring proper classifier performance the most important ones are those based on occurrence of grammatical structures. We also identified tree-based classifiers as most promising in terms of accuracy – for both training and testing datasets. In general, authors' classification for selected features and both instances proved to be reasonably accurate. It can be noted that presented approach can be used for different languages – provided that suitable parsing procedure could be conducted beforehand.

Further work in the research area of this contribution will involve employing genetic feature set selection. Supplementary experimental studies on effectiveness of various classifiers are planned as well. The possibility of using authorship attribution techniques for identifying source code creator constitutes likewise a promising area of upcoming research.

The study is co-funded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

References

[1] Argamon S., *Levitan S., Measuring the Usefulness of Function Words for Authorship Attribution*, Proc. Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, paper no 162, 2005.

[2] Broccias C., *Cognitive linguistic theories of grammar and grammar teaching*, [in:] De Knop S., De Rycker T. (eds.), *Cognitive Approaches to Pedagogical Grammar*, Walter de Gruyter, Berlin 2008, 67-90.

[3] Cheng N. Chandramouli R., Subbalakshmi K.P., *Author gender identification from text*, Digital Investigation, vol. 8, 2011, 78-88.

[4] Eder M., Rybicki J., *Do birds of a feather really flock together, or how to choose training samples for authorship attribution*, Literary and Linguistic Computing /to appear.

[5] Grieve J., *Quantitative authorship attribution: An evaluation of techniques*, Literary and Linguistic Computing, vol. 22, 2007, 251-270.

[6] Hastie T., Tibshirani R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York 2009.

[7] Hoover D.L., *Another Perspective on Vocabulary Richness*, Computers and the Humanities, vol. 37, 2003, 151-178.

[8] Jagadev A.K, Devi S., Mall R., *Soft Computing for Feature Selection*, [in:] Dehuri S., Cho, S.-B. (eds.), *Knowledge Mining using Intelligent Agents*, Imperial College Press, London 2011, 217-258.

[9] Jockers M.L., Witten D.M., *A comparative study of machine learning methods for authorship attribution*, Literary and Linguistic Computing, vol. 25, 2010, 215-223.

[10] Jurafsky D., Martin J.H., *Speech And Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, New Jersey 2009.
[11] Karlgren J., Eriksson G., *Authors, genre, and linguistic convention*, Proc. SIGIR Workshop on Plagiarism Analysis, Author ship Attribution, and Near-Duplicate Detection, 2007, 23-28.
[12] Klein D., Manning C.D., *Accurate Unlexicalized Parsing*, Proceedings of the 41st Meeting of the Association for Computational Linguistics, 2003, 423-430.
[13] Koppel M. Schler J., Argamon S., *Authorship attribution in the wild*, Language Resources and Evaluation, vol. 45, 2011, 83-94.
[14] Kowalski P.A., *Procedure of feature extraction from face image for biometrical system* (in Polish), Technical Transactions, vol. 1-AC/2012, Cracow University of Technology Press, 55-79.
[15] Layton R., *Authorship Attribution for Twitter in 140 Characters or Less*, Cracow University of Technology Press, Proc. 2nd Cybercrime and Trustworthy Computing Workshop, 2010, 1-8.
[16] Łukasik S., Haręza M., Kaczor M., *Grammatical structures ranking (supplementary material)*, http://www.pk.edu.pl/~szymonl/nauka/Author_suppl.pdf (date of access: 9.10.2013).
[17] Łukasik S., Kulczycki P., *Using topology preservation measures for high-dimensional data analysis in a reduced feature space* (in Polish), Technical Transactions, vol. 1-AC/2012, Cracow University of Technology Press, 5-15.
[18] Panicheva P., Cardiff J., Rosso P., *Personal Sense and Idiolect: Combining Authorship Attribution and Opinion Analysis*, [in:] Proc. International Conference on Language Resources and Evaluation, Valletta, paper no. 10.491, 2010.
[19] Penn Treebank Project, http://www.cis.upenn.edu/~treebank/
[20] Punctuation, [in:] Merriam-Webster's Collegiate Dictionary: Eleventh Edition, 1009, Merriam-Webster, Springfield 2004.
[21] Radford A., *Minimalist Syntax: Exploring the structure of English*, Cambridge University Press, Cambridge 2004.
[22] Stamatatos E., *A survey of modern authorship attribution methods*, Journal of the American Society for Information Science and Technology, vol. 60, 2009, 538-556.
[23] Wang L.P, Fu X.J., *Data Mining with Computational Intelligence*, Springer, Berlin 2005.
[24] Zheng R., Li J., Chen H., Huang Z., *A framework for authorship identification of online messages: Writing style features and classification techniques*, Journal of the American Society of Information Science and Technology, vol. 57, 2006, 378-393.