УДК 519.68; 681.513.7; 612.8.001.57; 007.51.52
DOI: 10.25559/SITITO.15.201903.572-578

Method for Author Attribution Using Word Embeddings

C. K. Simon[1*], I. V. Sochenkov[2,3]
[1] Peoples' Friendship University of Russia, Moscow, Russia
6 Miklukho-Maklaya Str., Moscow 117198, Russia
[*] conradsimon@hotmail.com
[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
44-2 Vavilov Str., Moscow 119333, Russia
[3] Lomonosov Moscow State University, Moscow, Russia
1, Leninskie gory, Moscow 119991, Russia

Abstract

In this paper we look at a methodology of revealing an unknown document's author through the use of extracting the author's characteristics from their writing style The method used explores identifying sources of unknown documents, using a model of distributive semantics to form a set of queries to a search engine. The dataset used is the PAN @ CLEF 2019 shared task on Cross-domain Authorship Attribution are in the following languages: English, French, Italian, and Spanish, each of which contains 5 problematic questions, which gives a total of 20 problematic questions. The problem relates to Natural Language Programming where the process is done through the attribution of the user that can be used to identify an author's work. The method explores identifying sources of unknown document, using a model of distributive semantics to form a set of queries to a search engine. The method used to reveal the unknown authors is done through distributional semantics; this is based on the following hypothesis: the linguistic units that are observed in close contexts have similar semantic meaning, in this area when looking at linguistics this is calculated based on the proximity of linguistic elements in terms of semantic load based on their distribution in large textual boxes.

Keywords: distributional semantics, source retrieval, anti-plagiarism, Author Attribution, Unknown Author, Natural language programming.

Funding: This article presents the research results of the project "Text mining tools for big data" as a part of the program supporting Technical Leadership Centers of the National Technological Initiative "Center for Big Data Storage and Processing" at the Moscow State University (Agreement with Fund supporting the NTI-projects No. 13/1251/2018 11.12.2018). The reported study is partially funded by the Russian Foundation for Basic Research (Project No. 18-29-03187) and supported by the "RUDN University Program 5-100".

For citation: Simon C.K., Sochenkov I.V. Method for Author Attribution Using Word Embeddings. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(3):572-578. DOI: 10.25559/SITITO.15.201903.572-578

Метод атрибуции автора с использованием Word Embeddings

С. К. Конрад[1*], И. В. Соченков[2,3]
[1] Российский университет дружбы народов, г. Москва, Россия
117198, Россия, г. Москва, ул. Миклухо-Маклая, д. 6
[*] conradsimon@hotmail.com
[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия
119333, Россия, г. Москва, ул. Вавилова, д. 44-2
[3] Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия
119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1

Аннотация

В этой статье рассматривается методология определения автора текста с помощью анализа стиля письма и определения особенностей, характерных для конкретного автора. Данный метод исследует возможности идентификации источников анализируемых документов с использованием модели дистрибутивной семантики для формирования набора запросов для поисковой машины. Используемый набор данных является совместной задачей PAN @ CLEF 2019 в Кросс-доменной Атрибуции Авторских прав на таких языках как английский, французский, итальянский и испанский, каждый из которых имеет 5 задач, что в совокупности ставит 20 задач. Общая задача, объединяющая эти 20 задач, связана с программированием на естественном языке, в рамках которого данный процесс осуществляется через атрибуцию пользователя, которая может быть использована для идентификации работы автора. Приведенный здесь метод исследует выявление источников неизвестного документа, используя модель дистрибутивной семантики для формирования набора запросов к поисковой системе. Метод, используемый для выявления неизвестных авторов, базируется на дистрибутивной семантике и на следующей гипотезе: лингвистические единицы, которые присутствуют в сходных контекстах, имеют сходное семантическое значение. Анализируемые лингвистические единицы рассчитываются, исходя из близости лингвистических элементов с точки зрения семантической нагрузки, основанной на их распределении в больших текстовых отрывках.

Ключевые слова: дистрибутивная семантика, поиск источника, антиплагиат, атрибуция авторства, неизвестный автор, программирование на естественном языке.

Финансирование: в данной статье представлены результаты исследования проекта «Инструменты текст майнинга для больших данных» в рамках программы поддержки Центров технического лидерства Национальной технологической инициативы по направлению «Центр хранения и анализа больших данных» при Московском государственном университете (соглашение с Фондом поддержки НТИ-проектов № 13/1251/2018 от 11.12.2018). Представленное исследование частично финансируется российским фондом фундаментальных исследований (проект № 1). 18-29-03187) и при поддержке «Университетской программы РУДН 5-100».

Для цитирования: *Конрад С. К., Соченков И. В.* Метод атрибуции автора с использованием Word Embeddings // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 3. С. 572-578. DOI: 10.25559/SITITO.15.201903.572-578

Introduction

In this research we will look into an interesting task, defining an author of the text based on attributes of the author which can be derived from the text, this done through the use of the training sample, in which there are several corps of documents of various authors and text of unknown authors whose authorship needs to be determined. The concept of the task bears some similarity to plagiarism detection where if the texts have high similarity, there is a high probability that this is the same text.

We developed a methodology to identify authors **from the PAN@ CLEF 2019 shared** task on Cross-domain Author Attribution in the languages English, French, Italian and Spanish each of which had five problems totaling twenty problems in total. What was explored in this work is Author Attribution which is a novel source of texts which has been targeted for analysis based on the writing characteristics of authors; the work being assessed is done through style-content dichotomy that is central to it such as a collection of fan fiction works for the task of cross-domain Authorship Attribution. The goal of this work is to develop a method for detecting Author Attribution based on distributive semantics; this is based on the following hypothesis: the linguistic units that are observed in close contexts have similar semantic meaning, in this area when looking at linguistics this is calculated based on the proximity of linguistic elements in terms of semantic load based on their distribution in large textual boxes [1]. Linguistic units can be words, phrases, and whole texts. This area allows you to perform such tasks as identifying the semantic proximity of words and phrases (building a synonym dictionary — thesaurus), defining the subject of the document, paraframe modeling, etc.

The following tasks were set: to

1. compose a mathematical model of distributive semantics;
2. develop an algorithm that, based on distributive semantics, will determine the degree of similarity of texts;
3. implement this algorithm in the form of a computer program;
4. run the program on the test set and get the results;
5. process the results and calculate the accuracy and recall of the algorithm;
6. analyze the results, provide recommendations for achieving the best result.

When processing on a computer, each document can be represented as a vector model, the text being considered as an unordered set of terms (a term is the word of which the text consists) [2].

All words from the text are being processed by being ordered; for a certain document you compile a list of all words in a sorted form, including those words that are not in this document but are found in other texts, then the document vector is obtained in the common vector space, while the dimension of the space is equal to the dimension of the vector and has the same value for all documents of the resulting vector space [3].

Related work

Closed-set authorship attribution is a task with rich relevant literature [4, 5]. 2017 and 2018 task of PAN@Clef Author Attribution included corresponding shared tasks [6, 7]. However, they only examined the case where both training and test documents belong to the same domain, as it is the case for the vast majority of published studies in this area. Crossdomain authorship attribution has been sporadically studied in the last decade [8-13]. In such cases, training and test texts belong to different domains that may refer to topic, genre, or language. To control a topic, usually general thematic categories are defined and all texts are pre-assigned to a topic. For example, Koppel et al. uses three thematic categories (ritual, business, and family) of religious Hebrew-Aramaic texts [14]. Newspaper articles are considered by Mikros and Argiri [15] (classified into two thematic areas: politics and culture) and Stamatatos (classified into four areas: politics, society, world, and UK) [11]. Another approach is to use a controlled corpus where some individuals are asked to write texts on a specific, well-defined topic [17].

Another important cross-domain perspective concerns cross-genre conditions. In general, it is hard to collect texts by several authors in different genres. Kestemont et al. make use of literary texts (theater plays and literary prose) [18] while Stamatatos explores differences between opinion articles and book reviews published in the same newspaper. Finally, a controlled corpus can be built, where each subject (author) is asked to write a text in a set of genres (e.g., email, blog, essay). The most extreme case concerns crosslanguage conditions where training and test texts are in different languages [19]. The focus went from clustering authors within documents [20] to the detection of positions where the style, i.e., the authorship, changes [21]. In general, all those tasks imply an intrinsic, stylometric analysis of the texts, as no reference corpora are available. In general, related work targeting multi-author documents is rare. While there exist several approaches for the related text segmentation problem, where a text is divided into distinct portions of different topics, only few approaches target segmentation by other criteria, especially not by authorship.

Methodology

Dataset

The accompanying JSON format[1] [22] files are available on the **PAN @ Clef** Cross-Domain Authorship Attribution 2019 are used as mapping data to identify information about the authors of the known text and the unknown text. These files contain a description of the directory structure, a list of folders for analysis, a list of texts for each author, a list of unknown texts and a file that stores information about the authors of unknown texts (for testing). The JSON files are as followed:

Collection-info.json which indicates a collection of such problems, that includes all relevant information such as the language giving in the format ‘en’ for English, ‘fr’ for French, ‘it’ for Italian and ‘sp’ for Spanish for each of the 20 problems given.This is located in the main folder of the task

An example of the collection-info.json file data is as follows:

```
[ { “problem-name”: “problem00001”,
 “language”: “en”,
“encoding”: “UTF-8” },
{ “problem-name”: “problem00002”,
 “language”: “fr”,
 “encoding”: “UTF-8” }, ... ]
```

Prolem-info.json which indicates the name of all authors in the form of candidate0000* which increments based on the amount of authors. The naming scheme of the folders located in problem0000* is based on the candidate numbers.

[1] Word vectors for 157 languages. fastText [Electronic resource]. Available at: https://fasttext.cc/docs/en/crawl-vectors.html (accessed 30.07.2019). (In Eng.)

An example of the problem-info.json file data is as follows:

```
{
 "unknown-folder": "unknown",
 "candidate-authors": [ { "author-name": "candidate00001" },
{ "author-name": "candidate00002" },... ] }
```

Ground-truth.json which indicates the true author of each unknown document, this file can be found in the main folder of each problem. Note that all unknown documents that are not written by any of the candidate authors belong to the <UNK> class.

An example of the groundtruth.json file data is as follows:

```
{
"ground_truth": [
{ "unknown-text": "unknown00001.txt",
"true-author": "candidate00002" },
{ "unknown-text": "unknown00002.txt",
"true-author": "<UNK>"}, ... ] }
```

Through the use of aforementioned JSON files and a mapping system we are able to use the candidate folders as training data; each candidate0000* folder which is located in each problem0000*.

The PAN @ Clef Cross-Domain Authorship Attribution 2019 comprises of **20** problems; **5** problems for each of the 4 languages (English , French, Italian and Spanish). Each of the 20 problem is made up of **9** candidates (authors), which has **7** texts for each author giving us a total of **5*9*7=315** training data for each language with a grand total of **1260** for all **4** languages.

4051 unknown was provided as testing data for all 4 languages; 1134 unknown documents in Spanish,1446 unknown documents in English, 920 unknown documents in French, 551 unknown documents.

Assessment

The files are assessed using vectors trained by Facebook using the FastText algorithm[2] [23] derived from data located on Wikipedia and Common Crawl[3]. These vector views are fairly recent (updated in 2018) and are provided in 157 languages which includes the languages needed for the task (English , French, Italian and Spanish).

Features to note:

*The words in the training did not undergo any processing, including stemming and translation to lower case. 300-dimensional vectors are provided for each language for the 2 million most popular words.

*The text vector was obtained by averaging the vectors of individual words.

Analysis and results

Optimizing parameters of Algorithms

The method of optimization used was Cross Validation and Grid Search, were used to optimize the parameters to create the best model for the Algorithms used [24]. A k-FOLD Cross Validation of 5 is used for all algorithms with a scoring parameter of "Accuracy". The grid search parameters for the KNeighbors are: 'n_neighbors': [***3,5,11,19,21***], 'weights': [***'uniform','distance'***], 'metric': [***'euclidean','manhattan'***]

and the parameters for LogisticRegression were '**solver**': [***'lbfgs', 'liblinear', 'sag', 'saga'***], 'C': [***1,10,100, 1000***], 'multi_class': [***'ovr', 'multinomial', 'auto'***], 'class_weight':[***'balanced'***].

Results

The mappings were done through the use of the JSON files which were provided by the PAN @ CLEF 2019 shared task on Cross-domain Author Attribution competition , we were able to successfully train the data using the candidate data, and the unknown authors were used as our testing data. To evaluate the performance of the algorithm, one should construct an inaccuracy matrix. Confusion Matrix (this is a type of binary classification [25] boundary which shows if the results are accepted or rejected) - a table that allows you to visualize the performance of the classification algorithm [26]. The matrix of inaccuracies is based on counting the number of times when the system made the right and wrong decision regarding the expected result. The matrix of inaccuracies is presented in Table 1

Table 1. Confusion Matrix

		The result	
		obtained The positive result	**The negative result The**
The expected result	**Expected positive result**	true-positive solution (TP: True Positive) *Must recognize - recognized the*	false-positive solution (FP: False Positive), **the first kind error** *Must recognize - did not recognize the*
	Expected negative result	false-negative solution (FN: False Negative), **error of the second kind** *Should not recognize - recognized a*	true-negative solution (TN: True Negative) *Should recognize - did not recognize*

At the same time, two types of errors are distinguished [27]:

False positive is a based on a coefficient of incorrect matches identifying that what we are searching for is there but was incorrectly classified, also known as type I error or the probability of false positive rate (FPR);

False Negative is a based on a coefficient of incorrect matches identifying that what we are searching for is not there but was incorrectly classified, also known as type II error or false non-recognition probability (FNR).

FNR can be calculated using the following formula:

$$FNR = \frac{M}{N}, \qquad (8)$$

where *m* is the number of failures, *N* is the total number of recognitions.

FPR is calculated by the formula:

$$FPR = \frac{k}{N}, \qquad (9)$$

where *k* is the number of incorrect recognitions, *N* is the total number of recognitions.

You can also select indicators such as precision, recall and F-measure.

The accuracy of the algorithm in terms of the inaccuracy table is calculated by the formula [28]:

$$precision = \frac{TP}{TP+FP}, \qquad (10)$$

[2] Common Crawl [Electronic resource]. Available at: https://commoncrawl.org (accessed 30.07.2019). (In Eng.)

[3] Common Crawl [Electronic resource]. Available at: https://commoncrawl.org (accessed 30.07.2019). (In Eng.)

Recall (the percentage of documents found by the algorithm belonging to a class relative to all documents of this class in the test set) can be calculated as follows [28]:

$$recall = \frac{TP}{TP + FN}, \qquad (11)$$

F_1-measure represents the harmonic average between precision and recall. It tends to zero if precision or recall tends to zero:

$$F_1 = \frac{2 * precision * recall}{recision + recall}, \qquad (12)$$

F1-score: 0.4659
Precision: 0.4924
Recall: 0.5115

Thus, the program recognizes texts with an accuracy of 49.24%, while 51.15% of those that are to be recognized are recognized correctly.

Conclusion

In this work, it was shown how using FastText algorithm to identify sources of unknown texts to identify authors from the PAN @ CLEF 2019 shared task on Cross-domain Author Attribution in the languages English , French, Italian and Spanish each of which had five problems totaling twenty problems. The text vector was obtained by averaging the vectors of individual word, The words in the training did not undergo any processing, including stemming and translation to lowercase. The following results were achieved in this work:

F1-score: 0.4659
Precision: 0.4924
Recall: 0.5115

The result achieved in this work was; F1-score of 0.4659, Precision: 0.4924, Recall:0.5115 the result fell slightly below the score of the baseline by F1-score: -0.1181, Precision: -0.1221, Recall: -0.0805 The precision suffering from the biggest loss.

The method of classifier used was KNeighbour and Logistic regression Classifier, the later receiving the better results. While the results of the PAN at CLEF 2019 shared task on Cross-domain Author Attribution is not currently available, in the previous year the PAN18-BASELINE was; F1-Score 0.584, precision:0.588, recall:0.692. In PAN at CLEF 2018 shared task on Cross-domain Author Attribution the work from the Custódio and Paraboni had the best overall performance the method used was char & word n-gram and a Neural Network classifier [29] with a result F1-Score 0.685, precision:0.672, recall:0.784 ..

Recommendations are given to achieve the best result

to reduce memory consumption by the program, you can use discharged matrices [30], but this can lead to an increase in the running time of the program;

the use of the base of frequently used words (in this work all words that are shorter than a predetermined number are cut off);

use of several types of binary classifiers to search for a more optimal boundary value.

use context-dependent word embeddings to predict the similarity of texts in that task

Acknowledgments: The authors are grateful to the reviewers for careful reading of the manuscript and helpful remarks.

Благодарности: авторы благодарят рецензентов за внимательное прочтение рукописи и полезные замечания.

References

[1] Rieger B.B. On Distributed Representations in Word Semantics. ICSI Technical Report TR-91-012. Berkeley, 1991. 29 pp. Available at: http://www.icsi.berkeley.edu/ftp/global/pub/techreports/1991/tr-91-012.pdf (accessed 30.07.2019). (In Eng.)

[2] Salton G., Wong A., Yang C.S. A vector space model for automatic indexing. *Communications of the ACM.* 1975; 18(11):613-620. (In Eng.) DOI: 10.1145/361219.361220

[3] Bondarchuk D.V. Vector space model of knowledge representation based on semantic relatedness. *Bulletin of the South Ural State University. Series "Computational Mathematics and Software Engineering".* 2017; 6(3):73-83. (In Russ., abstract in Eng.) DOI: 10.14529/cmse170305

[4] Neal T., Sundararajan K., Fatima A., Yan Y., Xiang Y., Woodard D. Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*. 2017; 50(6):86. 36 p. (In Eng.) DOI: 10.1145/3132039

[5] Stamatatos E. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology.* 2009; 60(3):538-556. (In Eng.) DOI: 10.1002/asi.21001

[6] Juola P. An overview of the traditional authorship attribution subtask. In: P. Forner, J. Karlgren, C. Womser-Hacker (Eds.) *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, 17-20 September, Rome, Italy, 2012. (In Eng.)

[7] Argamon S., Juola P. Overview of the international authorship identification competition at PAN-2011. In: V. Petras, P. Forner, P. Clough (Eds.) *Notebook Papers of CLEF 2011 Labs and Workshops*, 19-22 September, Amsterdam, Netherlands, 2011. (In Eng.)

[8] Sapkota U., Bethard S., Montes M., Solorio T. Not all character n-grams are created equal: A study in authorship attribution. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 2015, pp. 93-102. (In Eng.) DOI: 10.3115/v1/N15-1010

[9] Overdorf R., Greenstadt R. Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. *Proceedings on Privacy Enhancing Technologies.* 2016; 2016(3):155-171. (In Eng.) DOI: 10.1515/popets-2016-0021

[10] Sapkota U., Solorio T., Montes-y-Gómez M., Bethard S., Rosso P. Cross-topic authorship attribution: will out-of-topic data help? In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14).* 2014, pp. 1228-1237. Available at: https://www.aclweb.org/anthology/C14-1116.pdf (accessed 30.07.2019). (In Eng.)

[11] Stamatatos E. On the Robustness of Authorship Attribution Based on CharacterN-gram Features. *Journal of Law and Policy.* 2013; 21(2):421-439. (In Eng.)

[12] Sapkota U., Solorio T., Montes-y-Gómez M., Bethard S. Domain Adaptation for Authorship Attribution: Improved Structural Correspondence Learning. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Berlin, Germany, 2016; 1: 2226-2235. (In Eng.) DOI: 10.18653/v1/P16-1210

[13] Stamatatos E. Authorship attribution using text distortion.

In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.* Valencia, Spain, April 3-7, 2017. Association for Computational Linguistics, 2017; 1:1138-1149. Available at: https://www.aclweb.org/anthology/E17-1107.pdf (accessed 30.07.2019). (In Eng.)

[14] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research.* 2007; 8:1261-1276. Available at: http://www.jmlr.org/papers/volume8/koppel07a/koppel07a.pdf (accessed 30.07.2019). (In Eng.)

[15] Mikros G., Argiri E. Investigating Topic Influence in Authorship Attribution. In: B. Stein, M. Koppel, E. Stamatatos (Eds.) *Proceedings of the SIGIR'07 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07). Amsterdam, Netherlands, July 27, 2007. CEUR Workshop Proceedings.* 2007; 276:5. Available at: http://ceur-ws.org/Vol-276/paper5.pdf (accessed 30.07.2019). (In Eng.)

[16] Rocha A. et al. Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*. 2017; 12(1):5-33. (In Eng.) DOI: 10.1109/TIFS.2016.2603960

[17] Stamatatos E. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology.* 2018; 69(3):461-473. (In Eng.) DOI: 10.1002/asi.23968

[18] Kestemont M., Luyckx K., Daelemans W., Crombez T. Cross-Genre Authorship Verification Using Unmasking. *English Studies.* 2012; 93(3):340-356. (In Eng.) DOI: 10.1080/0013838X.2012.668793

[19] Bogdanova D., Lazaridou A. Cross-Language Authorship Attribution. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'2014)*, European Language Resources Association, Reykjavik, Iceland, 2014, pp. 2015-2020. (In Eng.)

[20] Stamatatos E., Tschuggnall M., Verhoeven B., Daelemans W., Specht G., Stein B., Potthast M. Clustering by Authorship Within and Across Documents. In: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.) *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. Workshop Proceedings.* 2016; 1609:691-715. Available at: http://ceur-ws.org/Vol-1609/16090691.pdf (accessed 30.07.2019). (In Eng.)

[21] Tschuggnall M., Stamatatos E., Verhoeven B., Daelemans W., Specht G., Stein B., Potthast M. Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering. In: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.) *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings*. 2017; 1866. Available at: http://ceur-ws.org/Vol-1866/invited_paper_3.pdf (accessed 30.07.2019). (In Eng.)

[22] Al-Shaikh A., Sleit A. Evaluating IndexedDB performance on web browsers. In: *2017 8th International Conference on Information Technology (ICIT)*, Amman, 2017, pp. 488-494. (In Eng.) DOI: 10.1109/ICITECH.2017.8080047

[23] Young J.C., Rusli A. Review and Visualization of Facebook's FastText Pretrained Word Vector Model. In: *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, Tokyo, Japan, 2019, pp. 1-6. (In Eng.) DOI: 10.1109/ICESI.2019.8863015

[24] Huang Q., Mao J., Liu Y. An improved grid search algorithm of SVR parameters optimization. In: *2012 IEEE 14th International Conference on Communication Technology*, Chengdu, 2012, pp. 1022-1026. (In Eng.) DOI: 10.1109/ICCT.2012.6511415

[25] Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004. (In Eng.)

[26] Fawcett T. An Introduction to ROC Analysis. *Pattern Recognition Letters.* 2006; 27(8):861-874. (In Eng.) DOI: 10.1016/j.patrec.2005.10.010

[27] Betz M.A., Gabriel K.R. Type IV Errors and Analysis of Simple Effects. *Journal of Educational Statistics.* 1978; 3(2):121-143. (In Eng.) DOI: 10.3102/10769986003002121

[28] Powers D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies.* 2011; 2(1):37-63. Available at: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf (accessed 30.07.2019). (In Eng.)

[29] Custódio J.E., Paraboni I. EACH-USP Ensemble cross-domain authorship attribution. In: L. Cappellato, N. Ferro, J.Y. Nie, L. Soulier (Eds.) *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings.* 2018; 2125:76. Available at: http://ceur-ws.org/Vol-2125/paper_76.pdf (accessed 30.07.2019). (In Eng.)

[30] Pissanetzky S. Sparse Matrix Technology. Academic Press, 1984. 336 p. (In Eng.)

*Submitted 30.07.2019; revised 15.08.2019;
published online 30.09.2019.*
*Поступила 30.07.2019; принята к публикации 15.08.2019;
опубликована онлайн 30.09.2019.*

About the authors:

Conrad K. Simon, Postgraduate Student of the Department of Information Technology, Faculty of Science, Peoples' Friendship University of Russia (6 Miklukho-Maklaya Str., Moscow 117198, Russia), ORCID: http://orcid.org/0000-0002-1309-1867, conradsimon@hotmail.com

Ilya V. Sochenkov, Head of the Department of Intelligent Technologies and System, Institute for Systems Analysis of Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (44-2 Vavilov Str., Moscow 119333, Russia); Principal Researcher, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russia), Ph.D. (Phys.-Math.), ORCID: http://orcid.org/0000-0003-3113-3765, sochenkov@isa.ru

All authors have read and approved the final manuscript.

Об авторах:

Симон Конрад Кеньон, аспирант кафедры информационных технологий, факультет физико-математических и естественных наук, 1 Российский университет дружбы народов (117198, Россия, г. Москва, ул. Миклухо-Маклая, д. 6),

ORCID: http://orcid.org/0000-0002-1309-1867, conradsimon@hotmail.com
Соченков Илья Владимирович, заведующий отделом интеллектуальных технологий и систем, Институт системного анализа РАН, Федеральный исследовательский центр «Информатика и управление» Российской академии наук (119333, Россия, г. Москва, ул. Вавилова, д. 44-2); руководитель направления по разработке методов искусственного интеллекта для анализа больших массивов текстовой информации в Центре хранения и анализа больших данных, Московский государственный университет имени М. В. Ломоносова (119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, ORCID: http://orcid.org/0000-0003-3113-3765, sochenkov@isa.ru

All authors have read and approved the final manuscript.