УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2016 Том 16 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2016 Vol. 16 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

FORENSIC LINGUISTICS: AUTOMATIC WEB AUTHOR IDENTIFICATION

A.A. Vorobeva[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: alice_w@mail.ru

Article info
Received 16.12.15, accepted 17.02.16
doi: 10.17586/2226-1494-2016-16-2-295-302
Article in English

For citation: Vorobeva A.A. Forensic linguistics: automatic web author identification. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 295–302, doi: 10.17586/2226-1494-2016-16-2-295-302

Abstract
Internet is anonymous, this allows posting under a false name, on behalf of others or simply anonymous. Thus, individuals, criminal or terrorist organizations can use Internet for criminal purposes; they hide their identity to avoid the prosecuting. Existing approaches and algorithms for author identification of web-posts on Russian language are not effective. The development of proven methods, technics and tools for author identification is extremely important and challenging task. In this work the algorithm and software for authorship identification of web-posts was developed. During the study the effectiveness of several classification and feature selection algorithms were tested. The algorithm includes some important steps: 1) Feature extraction; 2) Features discretization; 3) Feature selection with the most effective Relief-f algorithm (to find the best feature set with the most discriminating power for each set of candidate authors and maximize accuracy of author identification); 4) Author identification on model based on Random Forest algorithm. Random Forest and Relief-f algorithms are used to identify the author of a short text on Russian language for the first time. The important step of author attribution is data preprocessing - discretization of continuous features; earlier it was not applied to improve the efficiency of author identification. The software outputs top q authors with maximum probabilities of authorship. This approach is helpful for manual analysis in forensic linguistics, when developed tool is used to narrow the set of candidate authors. For experiments on 10 candidate authors, real author appeared in to top 3 in 90.02% cases, on first place real author appeared in 70.5% of cases.

Keywords
web author identification, authorship attribution, computational linguistics, information security

Acknowledgements
Materials were presented at the conference ISPIT-2015 on Information security and protection of information technology

УДК 004.056.5

КОМПЬЮТЕРНАЯ КРИМИНАЛИСТИКА: ИДЕНТИФИКАЦИЯ АВТОРА ИНТЕРНЕТ-ТЕКСТОВ

А.А. Воробьева[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: alice_w@mail.ru

Информация о статье
Поступила в редакцию 16.12.15, принята к печати 17.02.16
doi: 10.17586/2226-1494-2016-16-2-295-302
Язык статьи – английский

Ссылка для цитирования: Воробьева А.А. Компьютерная криминалистика: идентификация автора Интернет-текстов // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 2. С. 295–302. doi: 10.17586/2226-1494-2016-16-2-295-302

Аннотация
Интернет является уникальной системой с точки зрения его анонимности. Пользователь может получать доступ к созданию и распространению информации анонимно, т.е. вовсе без прохождения процедуры идентификации и аутентификации, либо имеет возможность создания неограниченного числа идентификаторов для распространения информации под вымышленными именами, или злоумышленник получает доступ к данным учетной записи пользователя и имеет возможность создания или распространения информации от чужого имени. Все это снижает качество обеспечения информационной безопасности. При информационном обмене в Интернете крайне важным является возможность идентифицировать или аутентифицировать пользователя, определить – является ли пользователь тем, за кого он себя выдает. Существующие методы идентификации пользователей – авторов коротких электронных сообщений на русском языке являются недостаточно эффективными. Возникает задача повышения точности идентификации и аутентификации пользователей – субъектов информационных процессов, размещающих некоторые электронные текстовые сообщения в сети Интернет. В данной работе предложен алгоритм идентификации автора Интернет-текстов, включающий следующие этапы: 1) извлечение идентификационных признаков автора; 2) дискретизация

непрерывных признаков; 3) отбор подмножества наиболее информативных признаков; 4) идентификация пользователя – автора сообщения (на основании алгоритма Random Forest). Ранее дискретизация непрерывных признаков в решении задачи идентификации пользователей (авторов сообщений) не применялась, однако она позволяет существенно повысить точность идентификации. Результатом работы являются q наиболее вероятных авторов текста. На основании предложенного алгоритма было разработано специализированное программное обеспечение. Проведенные эксперименты показали, что автор был абсолютно верно идентифицирован системой в 70,5% случаев, пользователь был идентифицирован системой в число трех наиболее вероятных в 90,02% случаев.

Ключевые слова
идентификация анонимных пользователей, идентификация автора, авторство сообщений, компьютерная лингвистика, информационная безопасность

Благодарности
Материалы представлены на конференции ISPIT-2015: Информационная безопасность и технологии защиты информации.

Introduction

Progress, rapid evolution and wide distribution of online communication tools (e.g. social networks, blogs, forums) has increased the dependence of the society on the information itself, its production, distribution and use. Internet is anonymous, this allows posting under a false name, on behalf of others (e.g, the name of any well-known person) or simply anonymous. Often the Internet is used by individuals for criminal purposes (such as anonymous threats, extremist statements, distribution of illegal materials or trade secrets) and by criminal or terrorist organizations as one of major communication channels [1, 2]. In this case, they will try to hide their identity to avoid the prosecuting (the device characteristics can be forged; one person can use multiple usernames within one or several sites). One of the basic concept of information security is authenticity that ensures that the identity of a subject is the identity claimed. In other words, authenticity is assurance that any exchange of information is from the source it claims to be from. Authenticity includes identification or the recognition of a name indicating a subject. Existing approaches and algorithms for author identification of web-posts on Russian language methods are not sufficiently effective on short texts of online communication. Therefore, development of proven methods, technics and tools for author identification is extremely important and challenging task. In this work was developed algorithm and software for automatic author identification of short web-posts on Russian language, using computational linguistics techniques.

Previous research

Authorship attribution task has long history, beginning from resolving the question of authorship of some Shakespeare sonnets and continues today with author identification of web-post. Writing style defines person like a fingerprint. Every person has unconscious writing habits, specific words, sentence and post structure, punctuations this all are the special markers and identification features of author. This gives us an opportunity to use some text features in automatic author identification [3–5].

In recent years, there were a lot studies on authorship attribution of online texts. In most of them SVM (Support Vector Machine) classification algorithm is used to classify texts to authors.

For author identification commonly are used different stylometric features (lexical, syntactic, structural, content-specific and so on): characters frequencies, N-gram frequencies, function words frequencies, vocabulary richness, word frequencies, words length and sentences length distribution, words collocations, sentences length, preferred word positions, prepositional phrase structure, parts of speech distribution, phrasal composition grammar etc [6–8]. Existing solutions have two main limitation:

- most of them are for Roman and German language groups [3–11];
- for Russian language authorship attribution was studied only for rather long texts. [12–14]. In [14] for texts on Russian language achieved accuracy of author identification (10 candidate authors, text length – 5000 characters) was 47%.

Author identification task

Given t_j – a web-post (or text) of unknown-authorship, a set of authors (candidate authors) $U = \{u_1, ..., u_k\}$ and set of their texts $T = \{t_1, ..., t_m\}$, where m – number of texts and k – is number of authors. So the author u_i can be presented as subset $T_j \in T$. It is assumed that the author of t_j is one of the U.

We have to find effective algorithm a: $t_j \rightarrow U$, that calculates the probability of authorship for each author to be an author of text t_j; then sort probabilities in descending order and select top q authors: P (u_q author t_j), $q < k$. In most existing approaches identification algorithm outputs only one author with P ($u_{\max}$ author t_j).

Number of authors q is calculated automatically for each set of candidates authors U and text t_j, as the closest values P (u_i author t_j) to P ($u_{\max}$ author t_j). This is one of the main differences between the proposed and existing approaches.

Features. Extraction, discretization and selection

In this work, three main types of features are extracted from texts: lexical, syntactic-structural, meta-features. Syntactic-structural group includes using of sentences with different structure and construction, frequencies of different punctuations, text decorations (bold, italic), the logical structure of the text (blocks, paragraphs), and others. Lexical group includes vocabulary richness, using of functional words and the specific expression, using of certain language constructs, using of abbreviations and acronyms, words in foreign languages, using of links, images, and others. Meta-features is some additional information of posts: day of the week and time of post. Full feature set contains 490 different features listed below (Table 1).

Syntactic-structural group (38 *features*)
– Frequency of each punctuation symbol: .,;:!?–\"; (9 features) – Frequency of each special symbols: @#$%^&*()=+{}'»«/|~`(20 features) – Total number of sentences – Frequency of links – Frequency of images – Frequency of paragraphs – Frequency of emphasis techniques: boldface font – Frequency of emphasis techniques: italic font – Frequency of emphasis techniques: boldface italic font – Frequency of emphasis techniques: underlines and strikethrough
Lexical group (450 *features*)
– Text length in characters – Frequency of uppers – Frequency of letters – Frequency of digits – Frequency of white spaces – Frequency of tab spaces – Frequency of all special symbols: @#$%^&*()=+{}'»«/|~` – Frequency of all punctuations: .,;:!?–\"; – Frequency of abbreviations: млн., руб., дол., евр., тыс., млрд., коп., см., т.д., т.п., пр., рис. – Frequency of character Ёё – Total number of words – Average word length – Frequency of short words length 1- 5 characters – Frequency of medium words length 6-10 characters – Frequency of long words length 11- 15 characters – Frequency of very long words length 16- 20 characters – Average sentence length in words – Average sentence length in characters – Frequency of short sentences length 1- 5 words – Frequency of short sentences length 6-12 words – Frequency of long sentences length more than 13 words – Frequency of words with various length in words (11 features) Frequency of function words (418 features)
Meta-features group (2 *features*)
– Post publication time (hour) – Day of the week

Table 1. List of extracted features

Continuous features discretization. The discretization is a part of the data preprocessing for some important reasons: building and validating of authorship identification model goes faster, discretization can provide some non-linear relations and it can harmonize heterogeneous data: some features are numerical and some are binary.

Experiments carried out earlier in this study showed that the use of the proposed approach – discretization of continuous features – could significantly improve the author identification accuracy.

Feature selection. Approach and algorithm. The aim of the feature selection is to find the best subset with maximum discriminative power. In all works on author identification the following approach was used features were selected to find the best subset for all authors, suitable for all identification tasks. However, in this work, to find the best feature subset with the most discriminating power and improve accuracy of author identification for text t_j, another approach was proposed: the best subset is selected for each identification task and, respectively, for each set of candidate authors.

To solve this task was applied Relief-f feature selection algorithm. Earlier was tested several popular feature selection algorithms, the best results showed two supervised feature selection algorithms: algorithm based on information gain ratio (GR) and Relief-f algorithm, described in [15, 16]. However, later it was discovered, that in the most author identification tasks Relief-f performs better.

Classification algorithms

Earlier was tested several popular classification algorithms: SVN, Naïve Bayes, Decision Tree (C4.5) and Random Forests (RF) [17]. The results of experiments are listed below (Table 2).

Classification algorithm	Accuracy
SVN	0.598245
Naïve Bayes	0.480865
Decision Tree	0.575125
Random Forest	0.659535

Table 2. Identification accuracy for different classification algorithms

All experiments showed that the best of all is Random Forest algorithm. Therefore, later for text-to-authors classification was used Random Forest algorithm.

Random Forest is an ensemble or set of decision trees. All decision trees are constructed by using a randomly selected subset of features from training data. Each decision tree of the ensemble classifies text to one of the authors. After RF is build it can predict author of new text t_j, it outputs class that is the mode of the classes of the individual trees, «wins» the author for which the highest number of trees «voted».

For the first time for the author identification of text on Russian language was used Random Forest algorithm [17], but there were several studies for other languages [8, 18].

Author identification

Full author identification task can be divided in two independent and parallel subtasks.

1. Collection of authors and texts.
2. Author identification of web-post.

Pre-stage. Collection of authors/texts, and features extraction. This is a very important step because it is necessary to have actual information about existing authors and changes in their writing-style. Therefore, actions of this stage have to be carried out continuously or in certain time intervals. Fig. 1 shows three main steps of this stage.

This stage starts with indexing and parsing web-pages of some website to collect data. After some new web-post is found, it is passing to the procedure of features extraction. Feature extractor component can analyze and extract features from the web-post, after that we have the vector representation of the web-post.

Then the author, his original web-posts and their vector representations are saved to the special Database.

Main stage. Author identification of web-post. Author identification algorithm include several steps or stages described below (Fig. 2).

If we have some web-post of unknown-authorship (t_j) and previously the expert selected several potential authors or candidate authors (U), then the process of author identification can be described as on the Figure below. The first step is to extract features from t_j and get the vector representations of all web-posts (T) of candidate authors U.

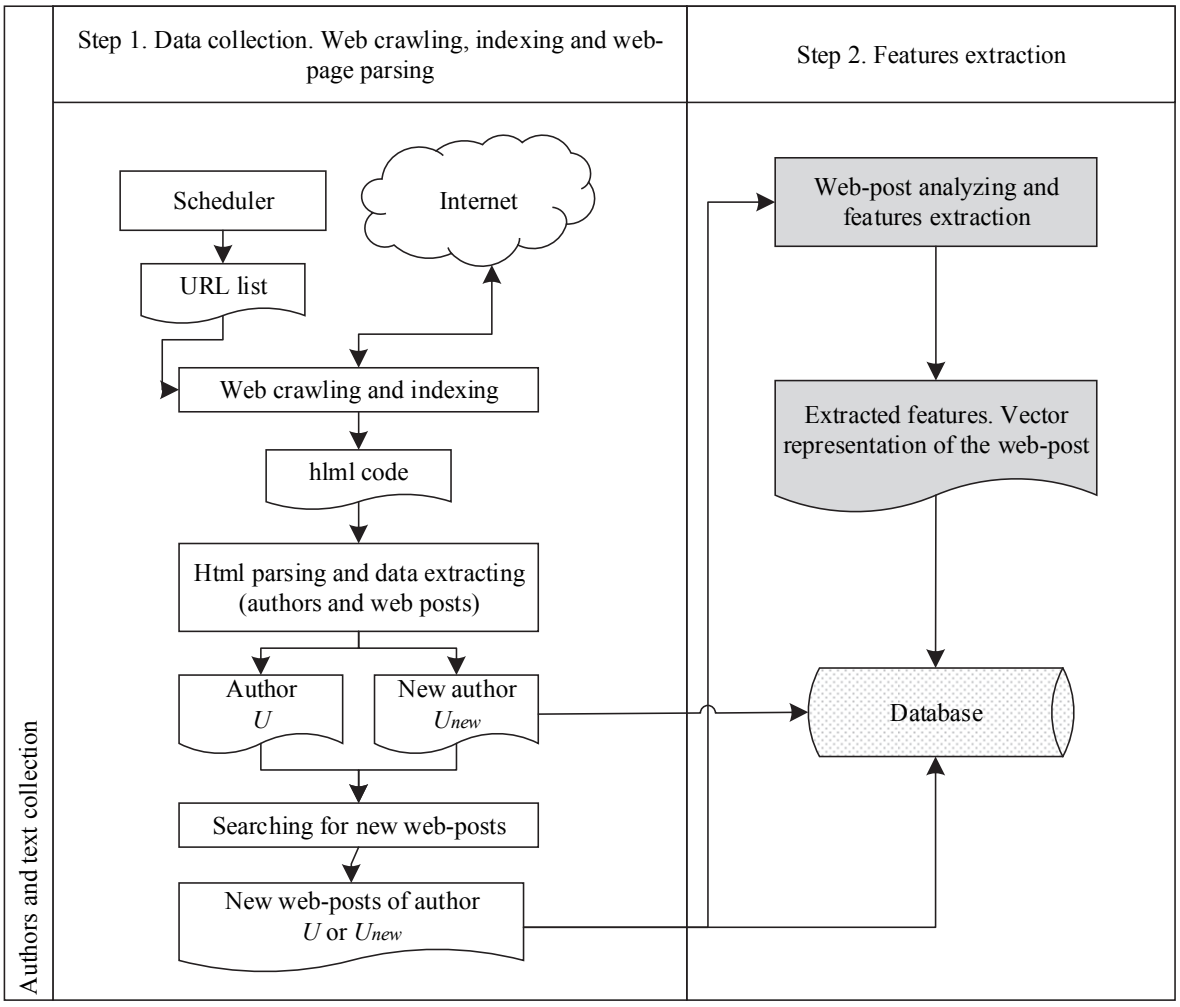

Fig. 1. Steps of the data collection and features extraction stage

After that, we are going to the features preprocessing. This process includes three main steps.

1. Discretization of continuous feature in a discrete feature constituted by a set of intervals of texts *T*.
2. Feature selection for *T*: Relief-f feature selection algorithm is used to find the best feature set (*F'*) with the most discriminating power for *U*. That is done to improve accuracy of author identification.
3. Features of t_j discretization and selection of *F'* subset.

Then, the author identification model (AIM) is build and validated. As author identification is the same task as classification text-to-authors, we train classification model on subset of texts and authors, then test it to validate the prediction power of model. The model is using *F'* features subset.

Validated authorship identification model can be used to identify author of text t_j.

The result of all previous steps is the list of top q authors from *U*, sorted by probabilities in descending order.

Experiments and results

Experiments were carried out on the text corpus contained 23546 web-post and 1004 authors. The text corpus was formed by collecting posts on Russian language from blog-hosting livejournal.com. The corpus is imbalanced and contains texts of different genres and topics. All texts have variable length, the maximum length of the texts was 5000 characters, and minimum text length is 100 characters, most of texts are from 100 to 499 characters length.

For the experiments were formed 80 sets of candidate authors (*U*), each included *l* authors, $l = \{2, 5, 10\}$, and their texts (20–25 texts for each author).

The accuracy *A* is ratio of correctly identified authors of texts T_c, to total number of test texts T_t (1).

$$A = \frac{T_c}{T_t} 100\% . \quad (1)$$

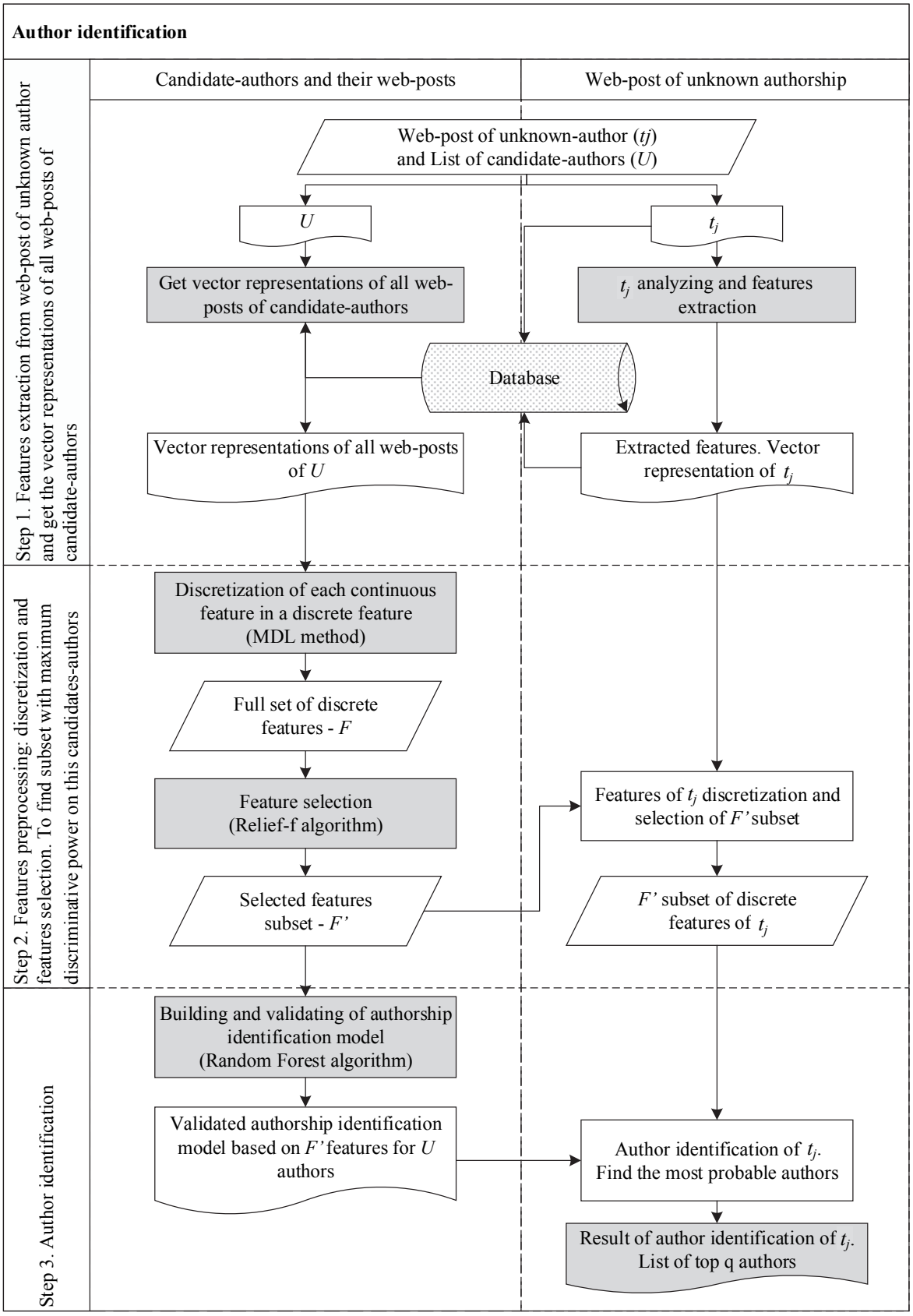

Fig. 2. Algorithm of author identification of web-post

Results are summarized on figures below, showing accuracy of authorship identification for different number of candidate authors. Experiments showed that the proposed approaches and developed tool could identify authors of web-posts on Russian language with satisfactory accuracy.

For experiments on 10 candidate authors, real author appeared in to top 3 in 90.02% cases, in top 2 in 84.1% cases, on first place real author appeared in 70.5% of cases, that is shown on Fig. 3. If classification algorithm returns up 3 top authors, it increases the accuracy of results returned to 90.02%.

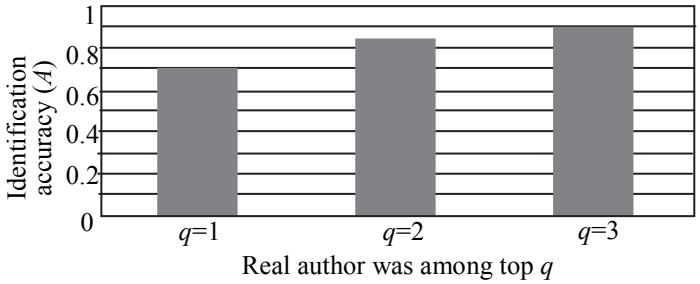

Fig. 3. Identification accuracies in cases real author appeared in top *q* authors

In particular, achieved accuracy for set of 10 candidate authors is 70.5%, for 5 authors – 89.3%, for 2 authors – 92.5%. The results are summarized in Fig. 4.

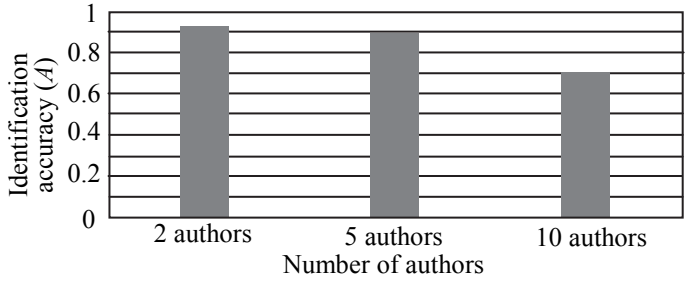

Fig. 4. Identification accuracies for different numbers of authors

Conclusion

In this work was developed the algorithm and software for author identification of web-posts, using computational linguistics techniques. The algorithm includes some important steps: 1) Feature extraction; 2) Features discretization (MDL method); 3) Feature selection with Relief-f algorithm (to find the best feature set with the most discriminating power for each set of candidate authors and maximize accuracy of author identification); 4) Author identification on model based on Random Forest algorithm. The result its work is top *q* authors with maximum probabilities of authorship.

To evaluate the accuracy some experiments on representative dataset were carried. Dataset was formed by collecting web-posts on Russian language from blog-hosting livejournal.com.

Experiments showed that the proposed approaches and developed software could identify authors of web-posts on Russian language with satisfactory accuracy.

Achieved accuracy of author identification of short texts (length less than 5000 characters) for set of 10 candidate authors is 70.5%, for 5 authors – 89.3%, for 2 authors – 92.5%, that is much better (about 23.5 %) than the accuracy obtained in previous researches [13].

References

1. Gvozdev A.V., Lebedev I.S. Model' analiza informatsionnykh vozmozhnostei v otkrytykh komp'yuternykh sistemakh. *Proc. VII Int. Conf. on Modern Problems of Applied Informatics*. St. Petersburg, 2011, pp. 45–47. (In Russian)
2. Vorobeva A.A. Analiz vozmozhnosti primeneniya razlichnykh lingvisticheskikh kharakteristik dlya identifikatsii avtora anonimnykh korotkikh soobshchenii v global'noi seti Internet. *Informatia i Kosmos*, 2014, no. 1, pp. 42–46. (In Russian)
3. Stamatatos E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 2009, vol. 60, no. 3, pp. 538–556. doi: 10.1002/asi.21001
4. Holmes D.I. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 1998, vol. 13, no. 3, pp. 111–117. doi: 10.1093/llc/13.3.111
5. Abbasi A., Chen H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 2005, vol. 20, no. 5, pp. 67–75. doi: 10.1109/MIS.2005.81
6. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification. *Lecture Notes in Computer Science*, 2006, vol. 4183, pp. 77–86.

7. Maitra P., Ghosh S., Das D. Authorship verification: an approach based on random forest. *Proc. 6th Conference and Labs of the Evaluation Forum, CLEF 2015*. Toulouse, France, 2015.
8. Pacheco M.L., Fernandes K., Porco A. Random forest with increased generalization: a universal background approach for authorship verification. *Proc. Conference and Labs of the Evaluation Forum, CLEF 2015*. Toulouse, France, 2015.
9. Afroz S. *Deception in Authorship Attribution. PhD thesis*. Drexel University, 2013, 104 p.
10. Haj Hassan F.I., Chaurasia M.A. N-gram based text author verification. *Proc. International Conference on Innovation and Information Management, ICIIM 2012*. Chengdu, China, 2012, vol. 36, pp. 67–71.
11. 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 for Information Science and Technology*, 2006, vol. 57, no. 3, pp. 378–393. doi: 10.1002/asi.20316
12. Fomenko V.P., Fomenko T.G. *Avtorskii invariant russkikh literaturnykh tekstov*. In Fomenko A.T. *Novaya Khronologiya Gretsii*. Moscow, 1995, vol. 2. (In Russian)
13. Khmelev D.V., Tweedie F.J. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 2001, vol. 16, no. 3, pp. 299–307. doi: 10.1093/llc/16.3.299
14. Romanov A.S. Metodika identifikatsii avtora teksta na osnove apparata opornykh vektorov. *Doklady TUSUR*, 2009, vol. 1, no. 2, pp. 36–42. (In Russian)
15. Kira K., Rendell L.A. A practical approach to feature selection. *Proc. 9th International Workshop on Machine Learning*, 1992, pp. 249–256. doi: 10.1016/B978-1-55860-247-2.50037-1
16. Kononenko I. Estimating attributes: analysis and extensions of RELIEF. *Lecture Notes in Computer Science*, 1994, vol. 784, pp. 171–182. doi: 10.1007/3-540-57868-4_57
17. Breiman L. Random forests. *Machine Learning*, 2001, vol. 45, no. 1, pp. 5–32. doi: 10.1023/A:1010933404324
18. Fatih Amasyali M., Diri B. Automatic Turkish text categorization in terms of author, genre and gender. *Lecture Notes in Computer Science*, 2006, vol. 3999, pp. 221–226.

Alisa A. Vorobeva – assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, alice_w@mail.ru

Воробьева Алиса Андреевна – ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alice_w@mail.ru