НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2017 Том 17 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2017 Vol. 17 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.89

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ИДЕНТИФИКАЦИИ ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ ПО КОРОТКИМ ЭЛЕКТРОННЫМ СООБЩЕНИЯМ

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 14.11.16, принята к печати 20.12.16
doi: 10.17586/2226-1494-2017-17-1-117-128
Язык статьи – русский

Ссылка для цитирования: Воробьева А.А. Отбор информативных признаков для идентификации Интернет-пользователей по коротким электронным сообщениям // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С. 117–128. doi: 10.17586/2226-1494-2017-17-1-117-128

Аннотация

Рассмотрена задача обеспечения идентификации и аутентификации субъектов информационных процессов, протекающих в среде Интернет и реализуемых с помощью коммуникационных средств Интернет-ресурсов по коротким электронным сообщениям (лингвистическая идентификация). Представлена комплексная многоуровневая модель Интернет-пользователя, включающая различные стилистические и лингвистические характеристики электронных сообщений. Сущность предлагаемого решения состоит в том, что из всех идентификационных признаков (лингвистических и стилистических характеристик) в каждой частной задаче идентификации предложено производить динамическое вычисление и отбор наиболее информативных признаков. Предлагаемое решение основано на том факте, что информативность идентификационных признаков отличается для различных пользователей и их групп. Расчет информативности и отбор признаков предложено производить на основе вычисления расстояния по значению признака до k ближайших соседей (алгоритм Relief-f). Проведены эксперименты по тестовым данным с различным количеством сообщений на одного пользователя. Результаты исследований показали, что использование динамического количества признаков, рассчитываемого для каждого набора пользователей, дает повышение точности идентификации в среднем на 4%, что почти на 1% выше, чем при использовании статического набора признаков. Предлагаемое решение наиболее эффективно при малом количестве сообщений одного пользователя.

Ключевые слова

идентификация Интернет-пользователей, лингвистическая идентификация, информационная безопасность

DYNAMIC FEATURE SELECTION FOR WEB USER IDENTIFICATION ON LINGUISTIC AND STYLISTIC FEATURES OF ONLINE TEXTS

A.A. Vorobeva[a]

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

Article info
Received 14.11.16, accepted 20.12.16
doi: 10.17586/2226-1494-2017-17-1-117-128
Article in Russian

For citation: Vorobeva A.A. Dynamic feature selection for web user identification on linguistic and stylistic features of online texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 117–128. doi: 10.17586/2226-1494-2017-17-1-117-128

Abstract

The paper deals with identification and authentication of web users participating in the Internet information processes (based on features of online texts). In digital forensics web user identification based on various linguistic features can be used to discover identity of individuals, criminals or terrorists using the Internet to commit cybercrimes. Internet could be used as a tool in different types of cybercrimes (fraud and identity theft, harassment and anonymous threats, terrorist or extremist statements, distribution of illegal content and information warfare). Linguistic identification of web users is a kind of biometric identification, it can be used to narrow down the suspects, identify a criminal and prosecute him. Feature set includes various linguistic and stylistic features extracted from online texts. We propose dynamic feature selection for each web user identification task. Selection is based on calculating Manhattan distance to k-nearest neighbors (Relief-f algorithm). This approach improves the identification accuracy and minimizes the number of features. Experiments were carried out on

several datasets with different level of class imbalance. Experiment results showed that features relevance varies in different set of web users (probable authors of some text); features selection for each set of web users improves identification accuracy by 4% at the average that is approximately 1% higher than with the use of static set of features. The proposed approach is most effective for a small number of training samples (messages) per user.

Keywords
web user identification, forensic linguistics, information security

Введение

Все современные исследователи подтверждают факт роста числа преступлений, совершаемых с использованием компьютеров и Интернет, таких как преследование, домогательства, вымогательство, анонимные угрозы, корпоративный шпионаж, терроризм и экстремизм[1]. Такие преступления совершаются легальными способами (через форумы, электронную почту, социальные сети, мессенджеры) и часто с использованием методов социальной инженерии. Рост их числа во многом обусловлен анонимной природой Интернета. Преступник может действовать скрытно, никаким образом не обнаруживая свою реальную личность. Гибкость идентичности и диссоциативная анонимность могут стимулировать преступное поведение в киберпространстве.

Для обеспечения безопасности в Интернет-пространстве и противодействия компьютерным преступлениям могут применяться методы идентификации и аутентификации пользователей, в том числе как часть систем разграничения доступа к Интернет-ресурсам или сервисам публикации электронных сообщений.

Одним из наиболее перспективных направлений развития технологий идентификации является биометрическая идентификация. В настоящей работе речь идет о такой ее разновидности, как лингвистическая идентификация Интернет-пользователей (идентификация по характерным свойствам и особенностям стиля письменной речи), т.е. по лингвистическим и стилистическим характеристикам электронных сообщений. Каждый человек имеет свой стиль письма, который составляет своеобразный уникальный «отпечаток» – набор характеристик, позволяющих его идентифицировать [1–3].

Лингвистическая идентификация Интернет-пользователей, помимо систем разграничения доступа, может применяться для решения задач в различных прикладных областях, например, в целях противодействия терроризму (установление того, что некоторое сообщение или заявление принадлежит некоторому террористу или экстремисту) [3], в уголовном праве (для определения автора анонимного сообщения, содержащего угрозу) [4], в гражданском праве (для разрешения споров об авторских правах), в компьютерной криминалистике (для определения автора вредоносного кода) [5], либо, в общем, для идентификации злоумышленника, отправляющего электронные сообщения [6, 7].

Обзор предыдущих исследований по лингвистической идентификации

Предыдущие исследования по лингвистической идентификации относятся к двум группам: это методы для идентификации автора литературного произведения и методы для идентификации пользователей по сообщениям, разработанные в основном для иностранных языков.

История идентификации автора литературного произведения начинается еще в XVIII веке; первыми внимание исследователей привлекли различные структурные характеристики текстов, а также любимые слова автора, специфические термины и выражения. Применение математических и статистических методов к анализу стиля (стилометрия) начинается только во второй половине XIX века. Несколько позже было доказано, что синтаксические особенности (такие как способы деления текста на параграфы, выделение прямой речи, сложность и длина предложений) сохраняются для различных текстов одного автора [8, 9].

Еще один подход к идентификации основан на анализе используемых слов и выявлении различных характеристик этих слов, так, в середине XIX века было определено, что частоты слов различной длины являются уникальной характеристикой автора, например, [10]. В ряде исследований были использованы методы идентификации, основанные на частотных характеристиках отдельных символов [11] и их последовательностей (униграмм, биграмм, *n*-грамм).

Частоты «функциональных» или служебных слов могут использоваться в качестве идентификационных признаков [12, 13]. В работе [14] было установлено, что с течением времени доля служебных слов в тексте остается постоянной.

Все исследования последнего десятилетия по лингвистической идентификации посвящены двум ключевым вопросам: поиску эффективного алгоритма идентификации (классификации) и выбору идентификационных признаков, обладающих наибольшей различающей способностью [15]. Особое внимание в настоящее время уделяется объективной оценке предлагаемых методов, а также сравнению различных

[1] Управление ООН по наркотикам и преступности. Всестороннее исследование проблемы киберпреступности 2013. URL: https://www.unodc.org/documents/organized-crime/cybercrime/Cybercrime_Study_Russian.pdf (дата обращения: 12.09.2016).

методов между собой. Существует достаточно большое количество работ в этой области, но почти все они анализируют тексты на английском языке, либо на языках германской группы.

Под точностью идентификации A понимается отношение количества правильно идентифицированных пользователей по сообщениям – $IdentU_{corr}$ к общему числу сообщений тестовой выборки – $|T_{tr}|$. Точность рассчитывается по формуле

$$A = \frac{IdentU_{corr}}{|T_{tr}|} \times 100\%.$$

Существует достаточно большое количество стилистических и лингвистических характеристик, которые могут выступать идентификационными признаками пользователя. В проанализированных работах используются различные признаки, однако в большинстве работ применяются n-граммы символов и слов [16], частоты служебных слов, частоты определенных частей речи, знаков пунктуации, слов и предложений определенной длины.

В работах [17] и [18] исследуется возможность идентификации пользователя по текстам электронных писем. Учитываются характеристики из пяти различных категорий: характеристики всего текста, частотные характеристики функциональных слов, плотность распределения длин слов, специфические характеристики электронных писем. Качество классификации существенно повышалось при совместном использовании специфических характеристик электронных писем с классическими характеристиками текста. В работе [19] доказано, что совместное использование характеристик различных типов позволяет повысить точность идентификации.

В работе [20] изучается влияние количества потенциальных пользователей и размера обучающей выборки на точность классификации. В работе доказывается, что с ростом количества пользователей точность идентификации уменьшается, при 145 потенциальных пользователях точность классификации составляет всего 12%.

Описанные исследования проводились для сообщений на различных иностранных языках, в них не проводилось исследование электронных сообщений на русском языке. На данный момент существует несколько работ, посвященных идентификации по сообщениям на русском языке, в первую очередь это работы [21–23].

Автором работы [21] была достигнута точность идентификации 98% при длине текста 20000 символов в результате обучения на трех примерах для одного пользователя. Разработанная автором методика основывается на применении метода опорных векторов, в качестве идентификационных признаков пользователя используются частоты букв русского языка, знаков пунктуации, наиболее частых триграмм символов и наиболее частых слов. При длине текста 5000 символов получена точность около 47%.

Сравнение проанализированных работ по идентификации пользователя по сообщениям на русском языке приведено в табл. 1.

Работа	Тип характеристик	Количество потенциальных пользователей	Длина сообщения и количество обучающих примеров	Точность идентификации A, %
[21]	Смешанные	10	более 20000 символов, 3 сообщения на 1 пользователя	98
[21]	Смешанные	10	5000 символов, 3 сообщения на 1 пользователя	47
[22]	Смешанные	250	–	Не оценивалась
[23]	Смешанные	более 10	500 слов, 1 сообщение на 1 пользователя	15
[23]	Смешанные	более 10	500 слов, 2 сообщения на 1 пользователя	19
[23]	Смешанные	более 10	500 слов, 9 сообщений на 1 пользователя	55

Таблица 1. Основные работы по идентификации пользователя

Ряд ограничений не позволяет применить существующие методы для идентификации пользователей по коротким электронным сообщениям на русском языке: это длина и язык текстов, для которых данные методы разрабатывались. Проанализированные методы лингвистической идентификации по коротким сообщениям на русском языке обладают достаточно низкой точностью.

В предыдущих исследованиях (табл. 1) предлагалось производить поиск и отбор лучших (или наиболее информативных) признаков по всем существующим пользователям и далее для представления пользователя использовать это минимальное количество идентификационных признаков. В данной работе предлагается использовать максимальное число идентификационных признаков, а отбор наиболее ин-

формативных производить для определенной группы потенциальных пользователей. Предлагаемый подход полезен, например, в случае идентификации в рамках одного Интернет-портала или при расследовании инцидентов нарушения информационной безопасности, когда некоторая конфиденциальная, внутренняя или порочащая репутацию информация распространяется анонимно в открытом доступе, а также при обнаружении на портале явления «астротурфинга». При идентификации устанавливается пользователь, с наибольшей вероятностью являющийся автором сообщения, т.е. для каждого пользователя определяется вероятность авторства. Далее полученные результаты могут быть обработаны в ручном либо автоматическом режиме в зависимости от решаемой задачи.

Модель представления Интернет-пользователя на основе стилистических и лингвистических характеристик электронных сообщений

В настоящей работе для идентификации используется подход с обучением на примерах, пользователь представляется как набор его сообщений, каждое из которых является обучающим примером. Подход позволяет комбинировать различные типы признаков (бинарные, дискретные, непрерывные), также в случае большого числа потенциальных пользователей такие методы обладают более высокой точностью [24].

Интернет-пользователю сопоставляется некоторый идентификатор – набор его электронных сообщений, каждое из которых, в свою очередь, представляется как набор лингвистических и стилистических характеристик (идентификационных признаков), представленных в векторной форме. Таким образом, электронное сообщение – вектор в пространстве признаковых описаний R^n, где n – это количество характеристик. Каждая из характеристик определяет некоторую особенность написания текста, внесенную подсознательно (например, пол лица, от которого ведется повествование, не может выступать в качестве признака пользователя).

При прохождении процедуры идентификации новое электронное сообщение преобразуется в набор характеристик, и производится сравнение с сообщениями, собранными ранее. Данное сравнение производится путем классификации, т.е. отнесения сообщения к определенному пользователю. Если ввести в рассмотрение n-мерное пространство признаков $\{Fi\}$, где $i = 1, \dots, n$, то каждое t_j сообщение (объект) в этом пространстве отображается точкой с координатами $\mathbf{t_j} = \mathbf{F_j} = (f_{j1}, \dots, f_{jn})$, а каждый класс объектов – пользователь $\mathbf{u_k}$ – множеством таких точек.

Анализ сообщений Интернет-пользователей позволил сделать вывод, что электронные сообщения, хотя и являются письменной речью, в реальности обладают многими особенностями устной [25]. Такая речь представляет собой сложную систему, в основе которой лежат коммуникативные, интеллектуальные, языковые навыки каждого конкретного пользователя. Многие из этих навыков проявляются в текстах в виде конкретных признаков, которые могут быть использованы при решении задач лингвистической идентификации, в то же время эти признаки индивидуализируют сами тексты.

Существует достаточно большое количество возможных характеристик электронных сообщений, которые могут быть использованы для лингвистической идентификации. Сгруппируем данные характеристики по их типу.

1. Лексические характеристики (**Fls** и **Flw**) могут быть, в свою очередь, подразделены на характеристики уровня слов (**Flw**) и характеристики уровня символов (**Fls**). В ходе исследования были отобраны характеристики текста, оказывающие положительное влияние на точность идентификации. В эту группу характеристик включены частоты служебных слов (в работе [14] было установлено, что с течением времени доля служебных слов остается постоянной). Частота использования буквы Ёё впервые используется в качестве идентификационного признака. Ранее, например, в работе [21] буквы Ёё замещались на Ее.
2. Синтаксические характеристики (**Fs**) традиционно объединяют различные особенности, определяющие структуру и конструкцию предложения. Общеизвестным является тот факт, что разным пользователям свойственно использовать разные конструкции предложений (например, простые и сложные предложения), кому-то свойственно использовать многоточия, а кому-то – тире. Если предыдущая группа характеристик объединяла в себе характеристики, выявляемые на уровне слов и символов, то синтаксические характеристики – это характеристики уровня предложений.
3. Структурные характеристики (**Fst**) отражают то, как пользователь представляет свой текст, т.е. то, как он разбивает его на предложения, на абзацы, использует ли ссылки, цитаты и изображения, какие виды начертаний текста он использует и как часто. Здесь выделяются две группы характеристик текста, которые можно использовать в качестве идентификационных признаков [26].
   – Форматирование текста сообщения. Ряд пользователей перед публикацией текста сообщения предпочитает определенным образом его оформлять: выделять шрифтом отдельные слова, словосочетания, предложения или абзацы.
   – Логическая структура текста документа. Даже самый короткий текст обычно подвергается логическому делению на абзацы, параграфы. Так как, например, абзац является текстовой единицей,

служащей для группировки однородных единиц изложения, исчерпывая один из его моментов (тематический, сюжетный и т.д.), он имеет композиционное, сюжетно-тематическое, ритмическое значение и связан со стилем пользователя.

4. Мета-характеристики текста (**Fm**) включают информацию, не имеющую прямого отношения к тексту сообщения. Но обычно в открытом доступе на исследуемом ресурсе существует запись о том, когда в сети появилось определенное сообщение. Как правило, пользователям свойственно проявление активности на веб-ресурсе, ведение разговоров с использованием средств электронной коммуникации и публикация сообщений в определенное время.

В табл. 2 приведен полный список характеристик электронных сообщений, используемых в данной работе. В работе используется уникальная комбинация стилистических, структурных, лексических характеристик и мета-характеристик сообщений, ряд которых для задачи идентификации Интернет-пользователя ранее не применялся.

Лексический уровень символов (**$F_j$ls**)	1. Длина сообщения в символах 2. Частота заглавных букв 3. Частота буквенных символов (Аа–Яя, Aa–Zz) 4. Частота цифр (0–9) 5. Частота пробелов 6. Частота символов табуляции 7. Частота появления буквы Ёё
Лексический уровень слов (**$F_j$lw**)	8. Общее число слов 9. Средняя длина слова 10–13. Частота слов определенной длины (4 признака) 14. Общее число предложений 15. Средняя длина предложения в словах 16. Средняя длина предложения в символах 17–19. Частота предложения определенной длины (3 признака) 20–39. Частота слов различной длины (20 признаков) 40–457. Частоты служебных слов (418 признаков) [27] 458. Частота использования аббревиатур (млн руб., дол., евр., тыс., млрд коп., см., т.д., т.п., пр., рис.)
Синтаксический уровень (**F_js**)	459. Частота всех знаков пунктуации (. , ; : ! ? – \ ”) 460–468. Частоты определенных знаков пунктуации [. , ; : ! ? – \ ”] (9 признаков) 469. Частота всех специальных символов (@#$%^&*()=+{}'<>/\|~`) 470–489. Частота встречаемости специальных символов [@#$%^&*()=+{}'<>/|~`] (20 признаков)
Структурный уровень (**F_jst**)	490. Частота использования гиперссылок 491. Частота использования изображений в сообщении 492. Частота использования переводов строк 493. Частота использования полужирного начертания текста 494. Частота использования курсивного начертания текста 495. Частота использования полужирного и курсивного начертания текста 496. Частота использования различных текстовых декораций – подчеркивания, зачеркивания
Мета-уровень (**F_jm**)	497. Час публикации сообщения 498. День недели публикации сообщения

Таблица 2. Характеристики электронного сообщения t_j

Комплексная многоуровневая модель представления Интернет-пользователя

Электронное сообщение обладает характеристиками, свойственными пользователю, его написавшему. Представляется возможным определить пользователя u_k как множество сообщений $\mathbf{T_k}$, где каждое сообщение $\mathbf{t_{kj}}$ представляется как кортеж, координатам которого сопоставляется множество характеристик, обусловленных особенностями используемой пользователем речи, $\mathbf{F_{kj}}$.

Разработана комплексная многоуровневая модель представления Интернет-пользователя (КММПП), описание используемых в ней идентификационных признаков приведено выше. В предлагаемой КММПП пользователь u_k представляется как множество его сообщений $\mathbf{T_k}$ или $\mathbf{F_k}$ – множество наборов векторных представлений сообщений пользователя u_k.

$$\mathbf{u_k} = \mathbf{T_k} = \{\mathbf{t_{k1}}, \dots, \mathbf{t_{kl}}\} = \mathbf{F_k} = \{\mathbf{F_{k1}}, \dots, \mathbf{F_{kl}}\},$$

где $\mathbf{u_k} \in \mathbf{U}$ и l – количество сообщений $\mathbf{u_k}$.

Так как, в свою очередь, сообщение $\mathbf{t_{kj}} \in \mathbf{T_k}$ представляется как вектор характеристик – $\mathbf{F_{kj}}$

$\mathbf{t_{kj}} = \mathbf{F_{kj}} = (\mathbf{F_{kj}ls}, \mathbf{F_{kj}lw}, \mathbf{F_{kj}s}, \mathbf{F_{kj}st}, \mathbf{F_{kj}m}) = (f_{kj1}, \dots, f_{kjn})$,

где n – количество характеристик, включенных в КММПП, n=498, $\mathbf{t_{kj}} \in \mathbf{T_k}$ **и** $\mathbf{F_{kj}} \in \mathbf{F_k}$, где $\mathbf{F_{kj}}$ – характеристики сообщения $\mathbf{t_{kj}}$, $\mathbf{F_{kj}ls}$ – лексические характеристики сообщения $\mathbf{t_{kj}}$ уровня символов, $\mathbf{F_{kj}lw}$ – лексические характеристики сообщения $\mathbf{t_{kj}}$ уровня слов, $\mathbf{F_{kj}s}$ – синтаксические характеристики сообщения $\mathbf{t_{kj}}$, $\mathbf{F_{kj}st}$ – структурные характеристики сообщения $\mathbf{t_{kj}}$, $\mathbf{F_{kj}m}$ – мета-характеристики сообщения $\mathbf{t_{kj}}$.

Следовательно u_k, можно представить в виде:

$$\mathbf{u_k} = \mathbf{T_k} = \{\mathbf{t_{k1}}, \dots, \mathbf{t_{kl}}\} = \begin{cases} (f_{11}, \dots, f_{1n}) \\ \dots \\ (f_{1l}, \dots, f_{1n}) \end{cases}.$$

КММПП пользователя u_k схематически представлена в табл. 3.

$\mathbf{u_k} = \mathbf{T_k}$	$\mathbf{T_k} = \{\mathbf{t_{k1}}, \dots, \mathbf{t_{kl}}\}$	$\mathbf{t_{kj}} = \mathbf{F_{kj}} = (\mathbf{F_{kj}ls}, \mathbf{F_{kj}lw}, \mathbf{F_{kj}s}, \mathbf{F_{kj}st}, \mathbf{F_{kj}m})$	$\mathbf{t_{kj}} = (f_{kj1}, \dots, f_{kjn})$
Пользователь $\mathbf{u_k}$	Сообщение $\mathbf{t_{k1}}$	Лексический уровень символов ($\mathbf{F_{k1}ls}$)	f_1 Длина сообщения
		Лексический уровень слов ($\mathbf{F_{k1}lw}$)	…
		Синтаксический уровень ($\mathbf{F_{k1}s}$)	
		Структурный уровень ($\mathbf{F_{k1}st}$)	
		Мета-уровень ($\mathbf{F_{k1}m}$)	f_n День недели публикации
	…	…	…
	Сообщение $\mathbf{t_{kl-1}}$	$\mathbf{F_{kl-1}}$	$(f_1, f_2, \dots, f_n)$
	Сообщение $\mathbf{t_{kl}}$	$\mathbf{F_{kl}}$	$(f_1, f_2, \dots, f_n)$

Таблица 3. Комплексная многоуровневая модель представления пользователя

КММПП позволяет производить идентификацию пользователей по характеристикам электронных текстовых сообщений длиной менее 5000 символов, что было ранее подтверждено экспериментально. Точность идентификации с использованием разработанной модели и различных современных методов классификации является более высокой по сравнению с результатами, полученными в предыдущих исследованиях (результаты экспериментов приведены в [28]).

Расчет информативности и отбор идентификационных признаков для формирования динамического стилистического профиля пользователя

Использования формальных методов идентификации пользователя по электронным сообщениям обладает одной существенной проблемой – выбор идентификационных признаков, обладающих высокой дискриминирующей способностью. Кроме того, с течением времени стиль написания сообщения пользователем может несколько изменяться, приобретать иные характерные особенности. При лингвистической идентификации необходимо это учитывать, т.е. отбор идентификационных признаков должен быть адаптивным.

В отличие от предыдущих работ (например, [21, 23, 29]) в модели представления пользователя предлагается использовать максимальное количество идентификационных признаков пользователя. При имеющемся количестве Интернет-пользователей и всем многообразии индивидуальных стилей невозможно предварительно исключить некоторые признаки, так как для некоторой частной задачи по идентификации пользователя и некотором наборе потенциальных пользователей именно эти признаки могут нести максимальную различающую способность, быть наиболее информативными.

Под информативностью или весом признака понимается величина, количественно характеризующая пригодность данного признака для распознавания пользователя, или мера, которая определяет, насколько хорошо данный признак разделяет сообщения различных пользователей и насколько плохо он отделяет сообщения одного пользователя между собой.

Задача идентификации пользователя по электронному сообщению – это задача многоклассовой классификации (имеется множество пользователей – классов). С целью повышения точности идентификации предлагается произвести уменьшение размерности вектора идентификационных признаков, т.е. уменьшение КММПП, что также приведет к уменьшению времени, необходимого для обучения классификатора.

Поиск и отбор подмножества идентификационных признаков, обладающих максимальной различающей способностью, предлагается производить именно для определенной группы потенциальных пользователей. На другой группе набор признаков будет совершенно иным, так как пользователи будут обладать другими отличительными особенностями стиля, что было подтверждено в ходе экспериментов, выполненных в настоящей работе. Данная задача может быть решена путем динамического вычисления и отбора наиболее информативных признаков из КММПП и формирования динамического стилистического профиля пользователя (ДСПП). На основании отобранных признаков будут сформированы динамические стилистические профили, что схематически можно представить в виде:

$$\textbf{КММПП } \mathbf{u_k} \quad \rightarrow \quad \textbf{ДСПП } \mathbf{u_k}$$

$$\mathbf{u_k} = \begin{cases} (f_{11}, \ \dots, \ f_{1n}) \\ \dots \\ (f_{l1}, \ \dots, \ f_{ln}) \end{cases} \quad \rightarrow \quad \mathbf{u_k} = \begin{cases} (f_{11}, \ \dots, \ f_{1m}) \\ \dots \\ (f_{l1}, \ \dots, \ f_{lm}) \end{cases}$$

Далее ДСПП будут использованы для лингвистической идентификации.

Задача отбора меньшего числа наиболее информативных признаков из исходного пространства признаковых описаний R^n или формирования пространства признаковых описаний R^m меньшей размерности, обладающих на данном наборе потенциальных пользователей наибольшей различающей способностью, может быть формализована следующим образом.

Пусть сообщение $\mathbf{t_{kj}} \in \mathbf{T}$ описывается n признаками, $\mathbf{F_{kj}} = (f_{kj1}, \dots, f_{kjn})$ – множество всех признаков (координат вектора), преобразованное признаковое пространство состоит из подмножества координат вектора:

$$\mathbf{F'_{kj}} = FSelect(\mathbf{F_{kj}}) = (f_{kj1}, \dots, f_{kjm}),$$

где $\mathbf{F'_{kj}} \in \mathbf{F_{kj}}$ u $m \leq n$; $\mathbf{F_{kj}}$ – исходное признаковое пространство для сообщения $\mathbf{t_{kj}}$; $\mathbf{F'_{kj}}$ – преобразованное признаковое пространство; $FSelect$ – функция преобразования.

Был проведен анализ применения метода формирования ДСПП и состава результирующих ДСПП для нескольких задач идентификации. Было установлено, что результирующие профили имеют различную длину, содержат различные признаки, и вес признака в каждой задаче отличается. Это позволяет сделать вывод, что информативность идентификационных признаков различается на разных наборах потенциальных пользователей.

Для формирования ДСПП и выбора оптимального количества информативных признаков отбор предлагается производить на основе вычисления расстояния по значению признака до k ближайших соседей (алгоритм Relief-f [30]). Алгоритм Relief-f ранее не применялся для отбора наиболее информативных признаков пользователей, однако он был достаточно эффективен при решении схожих задач в других предметных областях.

Отбор признаков производится на основе вычисления расстояния по значению признака до k ближайших соседей. В данной работе берется k=10, как рекомендовано в [30]. Каждому из признаков назначается определенный коэффициент, рассчитанный путем оценки расстояния до ближайших сообщений того же пользователя и до сообщений других пользователей.

Алгоритм отбора признаков включает следующую последовательность шагов:

1. веса для всех признаков $W[F] = 0$;
2. для всех сообщений набора $j = (1, l)$, где l – размер выборки:
   1. отбор случайного сообщения $\mathbf{t_j} \in \mathbf{T}, j = (1, l)$ пользователя u_x;
   2. поиск k ближайших сообщений того же пользователя – $\mathbf{H_c}$;
   3. для всех других пользователей $u \neq u_x$:
      – поиск k ближайших сообщений других пользователей – $\mathbf{M_c}(u), u \neq u_x$ (рис. 1);
   4. вычисление информативности для каждого признака $f_i, i = (1, n)$, где n – количество признаков в КММПП:

$$W[f_i] = W[f_i] - \sum_{c=1}^{k} \frac{diff(f_i, \mathbf{t_j}, \mathbf{H_c})}{l \times k} + \sum_{u \neq u_x} \sum_{c=1}^{k} \left[\frac{p(u)}{1 - p(u_k)} \times \frac{diff\left(f_i, \mathbf{t_j}, \mathbf{M_c}(u)\right)}{l \times k} \right], \quad (1)$$

$$diff(f, t_1, t_2) = \frac{|val(f, t_1) - val(f, t_2)|}{\max(f) - \min(f)}; \quad (2)$$

3. отбор признаков f, вес которых превышает заданное пороговое значение $\mathbf{F'_{kj}} = (f_{kj1}, \dots, f_{kjm})$, где $W[f_i] \geq \tau$, $\tau = 0{,}0$.

По формуле (1) вес признака тем больше, чем меньше расстояние по значению данного признака до сообщений того же пользователя и чем больше расстояние по значению данного признака до сообщений других пользователей. В данном алгоритме используется манхэттенское расстояние. Таким образом, вес тем выше, чем лучше он отделяет сообщения одного пользователя от сообщений других пользовате-

лей, и чем хуже разделяет сообщения пользователя между собой. Отбираются признаки с положительной информативностью, чей вес $W[f_i] \geq 0{,}0$.

При расчете весов в формуле (1) используется нормализованное расстояние, рассчитываемое по формуле (2). Также при расчете влияния признака на качество разделения сообщений пользователей между собой учитывается априорная вероятность появления сообщений данного пользователя. Если сообщений пользователя мало и вероятности не учитываются, то вес признака будет существенно увеличиваться на значение расстояния для пользователей, у которых мало сообщений.

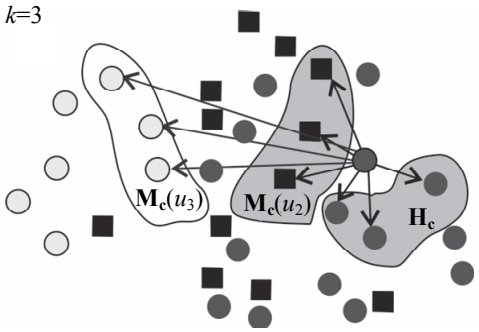

Рис. 1. Отбор признаков на основании расчета расстояния до *k* ближайших соседей с учетом априорных вероятностей

Эксперименты и оценка результатов

Для проверки эффективности разработанной модели представления пользователей и метода отбора информативных признаков при формировании динамического стилистического профиля пользователя была проведена серия экспериментов.

Корпус электронных сообщений был составлен из текстов, входящих в Национальный корпус русского языка, и дополнен текстами записей блогов на русском языке, содержащихся на портале LiveJournal, находящегося в публичном доступе. Корпус представляет собой коллекцию текстов различной тематики, для которых пользователь, являющийся автором сообщения, доподлинно известен. Отбор сообщений по тематике не производился. Распределение количества сообщений по длинам приведено на рис. 2. Данный корпус ранее использовался в работах [28, 31].

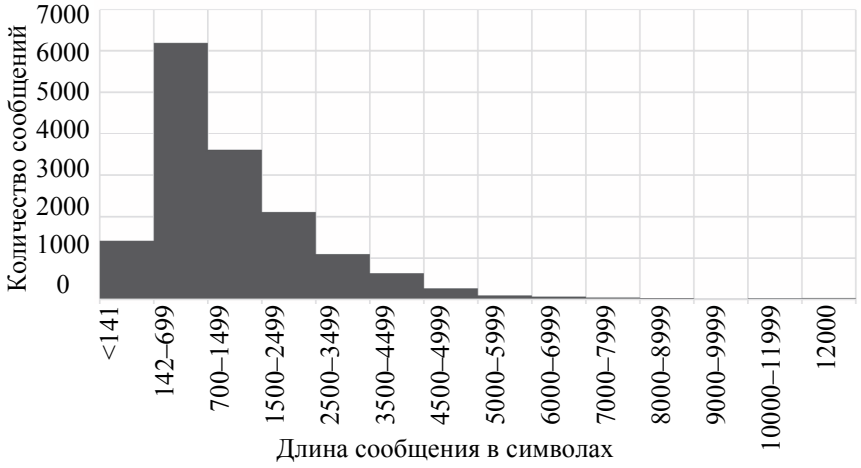

Рис. 2. Диаграмма распределений длин сообщений

Для сбора сообщений пользователей портала LiveJournal производилась индексация веб-страниц в автоматическом режиме с использованием разработанного поискового робота («веб-паук», краулер), который последовательно производит обход страниц пользователей и собирает сообщения, на них содержащиеся [32]. Информация о пользователях и сообщениях выделялась из содержимого веб-страниц по определенным критериям – последовательность html-тэгов, однозначно определяющая то, что в определенном месте html-кода содержится информация о пользователе или сообщении. При сборе сообщений из них удалялась вся незначимая информация (информация о разметке страницы, комментарии разработчиков страницы и пр.), иного дополнительного анализа или обработки сообщения не проводилось.

Из всех собранных сообщений производилось извлечение характеристик, приведенных в табл. 2. Большое число электронных сообщений делает ручной анализ крайне трудоемким, для автоматизации данного процесса было разработано специализированное приложение [33].

Помимо того, что Интернет-тексты являются достаточно краткими, существует вторая проблема – неравномерное распределение количества сообщений по пользователям. Недостаточное количество текстов обучающей выборки одного пользователя по сравнению с другими не должно снижать вероятность того, что данный пользователь будет верно идентифицирован. Для того чтобы приблизить эксперименты к реальной ситуации, были составлены наборы тестовых данных с различным уровнем несбалансированности (дисбалансом классов) и различным количеством сообщений на одного пользователя (табл. 4). В данные наборы были включены пользователи и сообщения, отобранные случайным образом. Подробное описание наборов было приведено в работе [28].

Уровень несбалансированности	Количество сообщений по пользователям (мин.:макс.)	
	Нормальное	Малое
Низкий	20:25	8:10
Средний	10:20	5:10
Высокий	5:25	2:10
Сбалансированный	24:25	10:10

Таблица 4. Наборы данных с различным уровнем несбалансированности

Для экспериментов были сформированы две группы наборов данных (с малым и нормальным количеством обучающих примеров), максимальное и минимальное количество текстов на одного пользователя приведено в табл. 4 (далее используются условные обозначения для указания на определенную группу наборов).

Все приведенные далее результаты были получены при использовании алгоритма классификации Random Forest (RF), так как в проведенных автором исследованиях было установлено, что именно он обладает максимальной точностью [28].

Было проведено сравнение точности идентификации при использовании динамического стилистического профиля пользователя и при использовании статического набора идентификационных признаков, включающего 100 наиболее информативных признаков [34]. Отбор наиболее информативных признаков для всех пользователей производился также с использованием приведенного выше алгоритма (Relief-f). Увеличение числа идентификационных признаков не оказывает положительного влияния на точность идентификации.

Оценка точности проводилась методом кросс-валидации по 8 блокам, соотношение обучающей и тестовой выборки – 90% и 10%. Результаты экспериментов приведены ниже (табл. 5, рис. 3).

Количество сообщений одного пользователя (мин.:макс.)	Точность идентификации A, %		
	Полная КММПП	Динамический отбор информативных признаков	Отбор наиболее информативных признаков для всех пользователей
10:10	69,56	74	72,59
25:25	75,42	78,17	77,09
8:10	66,72	71,72	70,32
20:25	75,33	78,03	76,63
5:10	62,21	67,66	66,88
10:20	71,5	75,01	74,65
2:10	60,52	65,79	64,2
5:25	70,21	73,76	74,35

Таблица 5. Точность идентификации при динамическом отборе информативных признаков по сравнению с базовой моделью и отбором информативных признаков по всем пользователям

Использование динамического количества признаков, рассчитываемого для каждого набора пользователей, дает улучшение точности идентификации в среднем на 0,93% по сравнению с использованием статического набора признаков и по сравнению с полным набором идентификационных признаков прирост составляет 4,08% (рис. 3). Наибольшее влияние динамический отбор оказывает в случаях, когда доступно малое количество текстов – средний прирост 1,3% и 5% соответственно. Наиболее подходящим для практического использования является динамический отбор идентификационных признаков для каждой задачи идентификации, так как он показал наиболее высокую точность для подавляющего большинства уровней несбалансированности и количества сообщений.

Результаты проведенных экспериментов подтверждают, что использование динамического количества и состава идентификационных признаков позволяет повысить точность идентификации и использовать в качестве идентификационных признаков значительно меньшее количество признаков, чем в исходном множестве.

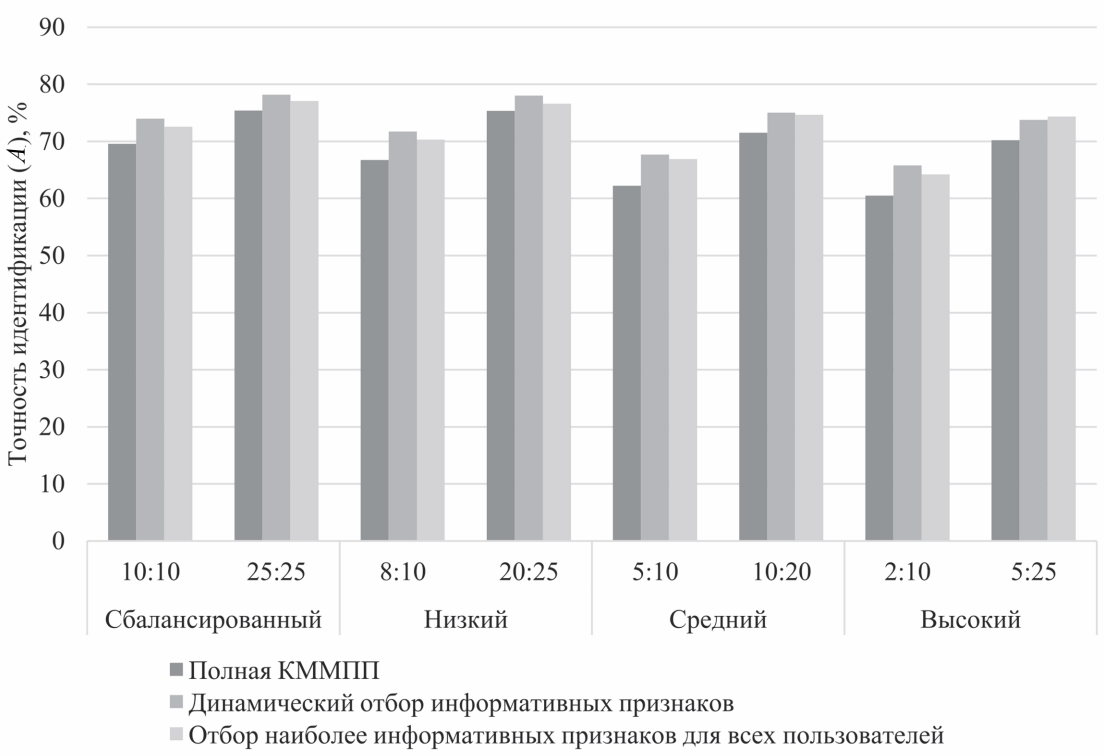

Рис. 3. Точность идентификации с использованием разработанного метода формирования динамических стилистических профилей пользователей

Заключение

Для представления пользователей при проведении лингвистической идентификации предлагается использовать максимальное количество признаков, так как при имеющемся количестве пользователей в сети Интернет и всем многообразии индивидуальных стилей невозможно предварительно исключить некоторые признаки. Также было установлено, что информативность (различающая способность) идентификационных признаков различается для разных пользователей.

В работе предлагается производить отбор и поиск подмножества идентификационных признаков каждой частной задачи идентификации, что позволяет повысить точность идентификации. Расчет информативности и отбор признаков производится на основе вычисления расстояния по значению признака до k ближайших соседей. Данный подход включен в метод формирования динамического стилистического профиля пользователя, который позволяет определить вес каждого из признаков и отобрать наиболее подходящие для каждой задачи идентификации.

Использование динамического количества признаков, рассчитываемого для каждого набора пользователей, дает улучшение точности идентификации в среднем на 4,08%, а по сравнению с использованием статического набора признаков прирост составляет 0,93% (рис. 3). Наибольшее влияние динамический отбор оказывает в случаях, когда доступно малое количество текстов – средний прирост 5% и 1,3% соответственно.

Литература

1. Лебедев И.С., Сухопаров М.Е. Методика идентификации авторства текстов коротких сообщений пользователей порталов сети интернет на основе методов математической лингвистики // В мире научных открытий. 2014. № 6.1 (54). С. 599–622.
2. Воробьева А.А., Гвоздев А.В. Идентификация анонимных пользователей Интернет порталов на основании технических и лингвистических характеристик пользователя // Научно-технический вестник механики и оптики. 2014. № 1(89). С. 139–144.
3. Abbasi A., Chen H. Applying authorship analysis to extremist-group web forum messages // IEEE Intelligent Systems. 2005. V. 20. N 5. P. 67–75. doi: 10.1109/MIS.2005.81
4. Frommholz I., al-Khateeb H.M., Potthast M., Ghasem Z., Shukla M., Short E. On textual analysis and machine learning for cyberstalking detection // Datenbank-Spektrum. 2016. V. 16.

References

1. Lebedev I.S., Sukhoparov M.Y. Methodologies of Internet portals users' short messages texts authorship identification based on the methods of mathematical linguistics. *In the World of Scientific Discoveries*, 2014, no. 6.1, pp. 599–622. (In Russian).
2. Vorob'yeva A.A., Gvozdev A.V. Anonymous website user identification based on combined feature set (writing style and technical features). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2014, no. 1, pp. 139–144. (In Russian).
3. 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
4. Frommholz I., al-Khateeb H.M., Potthast M., Ghasem Z., Shukla M., Short E. On textual analysis and machine learning for cyberstalking detection. *Datenbank-Spektrum*, 2016, vol.

N 2. P. 127–135. doi: 10.1007/s13222-016-0221-x
5. Rosenblum N., Zhu X., Miller B.P. Who wrote this code? Identifying the authors of program binaries // Lecture Notes in Computer Science. 2011. V. 6879. P. 172–189. doi: 10.1007/978-3-642-23822-2_10
6. Iqbal F., Binsalleeh H., Fung B.C.M., Debbabi M. A unified data mining solution for authorship analysis in anonymous textual communications // Information Sciences. 2013. V. 231. P. 98–112. doi: 10.1016/j.ins.2011.03.006
7. van der Knaap L., Grootjen F.A. Author identification in chatlogs using formal concept analysis // Proc. 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC2007). 2007. P. 181–188.
8. Yule G.U. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship // Biometrika. 1939. V. 30. N 3/4. P. 363–390. doi: 10.2307/2332655
9. Williams C.B. A note on the statistical analysis of sentence-length as a criterion of literary style // Biometrika. 1940. V. 31. N 3/4. P. 356–361. doi: 10.2307/2332615
10. Mendenhall T.C. A mechanical solution of a literary problem // Popular Science Monthly. 1901. V. 60.
11. Greg W.W., Yule G.U. The statistical study of literary vocabulary // The Modern Language Review. 1944. V. 39. N 3. P. 291. doi: 10.2307/3717870
12. Морозов Н.А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд // Известия Отдела русского языка и словесности Императорской Академии наук. 1915. Т. 20(7). С. 93–127.
13. Mosteller F., Wallace D. Inference and Disputed Authorship: The Federalist. Addison-Wesley, 1964. 287 p.
14. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов / В кн. Фоменко А.Т. Новая хронология Греции. Т. 2. М.: МГУ, 1995.
15. Potthast M., Braun S., Buz T., Duffhauss F., Friedrich F. et al. Who wrote the web? Revisiting influential author identification research applicable to information retrieval // Lecture Notes in Computer Science. 2016. V. 9626. P. 393–407. doi: 10.1007/978-3-319-30671-1_29
16. Haj Hassan F.I., Chaurasia M.A. N-gram based text author verification // Proc. Int. Conf. on Innovation and Information Management (ICIIM). Chengdu, China, 2012. V. 36. P. 67–71.
17. Corney M., Anderson A., Mohay G., de Vel. O. Identifying the authors of suspect email. 2001. Режим доступа: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf (дата обращения: 22.07.2016).
18. de Vel O., Anderson A., Corney M., Mohay G. Mining e-mail content for author identification forensics // ACM SIGMOD Record. 2001. V. 30. N 4. P. 55–64. doi: 10.1145/604264.604272
19. Zheng R., Li J., Huang Z., Chen H. 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. V. 57. N 3. P. 378–393. doi: 10.1002/asi.20316
20. Luyckx K., Daelemans W. Personae, a corpus for author and personality prediction from text // Proc. LREC. 2008. V. L08-1. P. 2981–2987.
21. Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста: автореф. … дисс.. канд. тех. наук. Томск, 2010. 26 с.
22. Сухопаров М.Е. Методика идентификации пользователей порталов сети интернет на основе методов математической лингвистики: автореф. … дисс. канд. тех. наук. СПб., 2015. 18 с.
23. Afroz S. Deception in Authorship Attribution. PhD thesis. Drexel University, 2013.
24. Yang M., Chow K.P. Authorship attribution for forensic investigation with thousands of authors // Proc. 29th IFIP Advances in Information and Communication Technology. 2014. V. 428. P. 339–350. doi: 10.1007/978-3-642-55415-5_28
25. Кузнецов А.В. Письменная разговорная речь в онлайн-коммуникации // Молодой ученый. 2011. № 3–2. С. 24–26.
26. Сигачёв А.С. Модель текста в виде набора числовых признаков // Интеллектуальные технологии и системы.

16, no. 2, pp. 127–135. doi: 10.1007/s13222-016-0221-x
5. Rosenblum N., Zhu X., Miller B.P. Who wrote this code? Identifying the authors of program binaries. *Lecture Notes in Computer Science*, 2011, vol. 6879, pp. 172–189. doi: 10.1007/978-3-642-23822-2_10
6. Iqbal F., Binsalleeh H., Fung B.C.M., Debbabi M. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 2013, vol. 231, pp. 98–112. doi: 10.1016/j.ins.2011.03.006
7. van der Knaap L., Grootjen F.A. Author identification in chatlogs using formal concept analysis. *Proc. 19th Belgian-Dutch Conference on Artificial Intelligence, BNAIC,* 2007, pp. 181–188.
8. Yule G.U. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, 1939, vol. 30, no. 3/4, pp. 363–390. doi: 10.2307/2332655
9. Williams C.B. A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, 1940, vol. 31, no. 3/4, pp. 356–361. doi: 10.2307/2332615
10. Mendenhall T.C. A mechanical solution of a literary problem. *Popular Science Monthly*, 1901, vol. 60.
11. Greg W.W., Yule G.U. The statistical study of literary vocabulary. *The Modern Language Review*, 1944, vol. 39, no. 3, pp. 291. doi: 10.2307/3717870
12. Morozov N.A. Linguistic spectra: Means to distinguish plagiarism from the true works of one or the other well-known author: Stilemetrichesky sketch. *Izvestiya Otdela Russkogo Yazyka i Slovesnosti Imperatorskoi Akademii Nauk*, 1915, vol. 20, pp. 93–127. (In Russian)
13. Mosteller F., Wallace D. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964, 287 p.
14. Fomenko V.P., Fomenko T.G. Avtorskii invariant russkikh literaturnykh tekstov. In Fomenko A.T. *Novaya Khronologiya Gretsii*. Moscow, MSU Publ., 1995, vol. 2. (In Russian)
15. Potthast M., Braun S., Buz T., Duffhauss F., Friedrich F. et al. Who wrote the web? Revisiting influential author identification research applicable to information retrieval. *Lecture Notes in Computer Science*, 2016, vol. 9626, pp. 393–407. doi: 10.1007/978-3-319-30671-1_29
16. Haj Hassan F.I., Chaurasia M.A. N-gram based text author verification. *Proc. Int. Conf. on Innovation and Information Management, ICIIM 2012*. Chengdu, China, 2012, vol. 36, pp. 67–71.
17. Corney M., Anderson A., Mohay G., de Vel. O. *Identifying the authors of suspect email*. 2001. Available at: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf (accessed: 22.07.2016).
18. de Vel O., Anderson A., Corney M., Mohay G. Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 2001, vol. 30, no. 4, pp. 55–64. doi: 10.1145/604264.604272
19. Zheng R., Li J., Huang Z., Chen H. 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
20. Luyckx K., Daelemans W. Personae, a corpus for author and personality prediction from text. *Proc. LREC*, 2008, vol. L08-1, pp. 2981–2987.
21. Romanov A.S. *Technique and Software Package for the Identification of Author of an Unknown Text*. PhD Thesis Eng. Sci. Tomsk, 2010, 26 p. (In Russian)
22. Sukhoparov M.E. *Technique for Identification of Internet Portals Users Based on Mathematical Linguistics Methods*. PhD Thesis Eng. Sci. St. Petersburg, 2015, 18 p. (In Russian)
23. Afroz S. *Deception in Authorship Attribution*. PhD thesis. Drexel University, 2013.
24. Yang M., Chow K.P. Authorship attribution for forensic investigation with thousands of authors. *Proc. 29th IFIP Advances in Information and Communication Technology*, 2014, vol. 428, pp. 339–350. doi: 10.1007/978-3-642-55415-5_28
25. Kuznetsov A.V. Written colloquial speach in online communication. *Molodoi Uchenyi*, 2011, no. 3-2, pp. 24–26. (In Russian).

2006. №7.
27. Vorobeva A.A. List of functional words used for web user (author) identification. 2016. doi: 10.13140/RG.2.2.30776.14080
28. Vorobeva A.A. Examining the performance of classification algorithms for imbalanced data sets in web author identification // Proc. 18th Conference of Open Innovations Association. 2016. P. 385–390. doi: 10.1109/fruct-ispit.2016.7561554
29. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification // Lecture Notes in Computer Science. 2006. V. 4183. P. 77–86. doi: 10.1007/11861461_10
30. Kononenko I. Estimating attributes: analysis and extensions of RELIEF // Lecture Notes in Computer Science. 1994. V. 784. P. 171–182. doi: 10.1007/3-540-57868-4_57
31. Vorobeva A.A. Forensic linguistics: automatic web author identification // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2016. V. 16. N 2. P. 295–302. doi: 10.17586/2226-1494-2016-16-2-295-302
32. Воробьева А.А., Пантюхин И.С., Швед Д.В. Средство для создания баз данных сообщений пользователей порталов сети Интернет. Свидетельство о регистрации программ для ЭВМ №2013661841. Опубл. 17.12.2013.
33. Воробьева А.А., Пантюхин И.С., Швед Д.В. Программный компонент лингвистического анализа и обработки текста для идентификации автора. Свидетельство о регистрации программы для ЭВМ №2014611567. Опубл. 5.02.2014.
34. Vorobeva A.A. 100 most informative features. 2016. Режим доступа: https://www.researchgate.net/publication/311510278_100_Most_informative_features (дата обращения: 08.12.2016). doi: 10.13140/RG.2.2.10289.58724

26. Sigachev A.S. Model of text as a set of numerical signs. *Intellektual'nye Tekhnologii i Sistemy*, 2006, no. 7. (In Russian).
27. Vorobeva A.A. *List of functional words used for web user (author) identification,* 2016. doi: 10.13140/RG.2.2.30776.14080
28. Vorobeva A.A. Examining the performance of classification algorithms for imbalanced data sets in web author identification. *Proc. 18th Conference of Open Innovations Association*, 2016, pp. 385–390. doi: 10.1109/fruct-ispit.2016.7561554
29. Houvardas J., Stamatatos E. N-gram feature selection for authorship identification. *Lecture Notes in Computer Science*, 2006, vol. 4183, pp. 77–86. doi: 10.1007/11861461_10
30. 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
31. 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
32. Vorob'eva A.A., Pantyukhin I.S., Shved D.V. *Tool to Create Database of Users Messages of the Internet Portals*. Certificate of State Registration of Computer Programs, no. 2013661841.
33. Vorob'eva A.A., Pantyukhin I.S., Shved D.V. *Software Component of Linguistic Analysis and Text Processing for Author Identification*. Certificate of State Registration of Computer Programs, no. 2014611567.
34. Vorobeva A.A. 100 most informative features. 2016. Available at: https://www.researchgate.net/publication/311510278_100_Most_informative_features (accessed: 08.12.2016). doi: 10.13140/RG.2.2.10289.58724

Авторы

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

Authors

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