14. Shcheglov K.A., Shcheglov A.Yu. Sistema zashchity ot zapuska vredonosnykh program [Malware startup protection system]. *Vestnik komp'yuternykh i informatsionnykh tekhnologii*, 2013, no. 5, pp. 38–43.
15. Shcheglov K.A., Shcheglov A.Yu. Realizatsiya metoda mandatnogo dostupa k sozdavaemym failovym ob"ektam sistemy [Implementation of mandatory access control to newly created file objects method]. *Voprosy zashchity informatsii*, 2013, no. 4 (103), pp. 15–20.

Щеглов Константин Андреевич – студент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; менеджер по развитию, ЗАО «НПП «Информационные технологии в бизнесе», Санкт-Петербург, Россия, schegl_70@mail.ru

Щеглов Андрей Юрьевич – доктор технических наук, профессор, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; генеральный директор, ЗАО «НПП «Информационные технологии в бизнесе», Санкт-Петербург, Россия, info@npp-itb.spb.ru

Konstantin Shcheglov Student, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics; development manager, JSC "Information Technologies in Business", Saint Petersburg, Russia, Scheglov.konstantin@gmail.ru

Andrei Shcheglov D.Sc., Professor, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics; General director, JSC "Information Technologies in Business", Saint Petersburg, Russia, info@npp-itb.spb.ru

УДК 004.89

ИДЕНТИФИКАЦИЯ АНОНИМНЫХ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-ПОРТАЛОВ НА ОСНОВАНИИ ТЕХНИЧЕСКИХ И ЛИНГВИСТИЧЕСКИХ ХАРАКТЕРИСТИК ПОЛЬЗОВАТЕЛЯ[1]

А.А. Воробьева[a], А.В. Гвоздев[a]

[a] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, Alice_w@mail.ru

Задача идентификации анонимных пользователей Интернет-порталов становится все более актуальной научной задачей, это обусловлено ростом числа интернет-пользователей, в том числе анонимных, ростом числа случаев совершения противоправных действий (например, анонимных угроз и экстремистских высказываний) и несовершенством существующих подходов и алгоритмов идентификации анонимных пользователей.

В контексте работы под идентификацией пользователя понимается распознавание анонимного пользователя в Интернете [1–5]. Распознавание производится путем соотнесения набора характеристик анонимного пользователя с характеристиками, собранными ранее и уже имеющимися в базе данных. К характеристикам пользователя относятся ряд технических (IP-адрес, версия операционной системы и пр.) и лингвистических (стиль письменной речи автора сообщения) характеристик. В работе рассматривается возможность идентификации пользователей по различным наборам таких характеристик (техническим, лингвистическим и комбинированным). Анализируется возможность применения различных методов классификации (метод опорных векторов, нейросети, логическая регрессия) для решения задачи по идентификации анонимных пользователей.

Проведенные эксперименты показали, что использование лингвистических характеристик совместно с техническими позволяет повысить точность идентификации анонимного пользователя Интернет-портала.

Ключевые слова: идентификация анонимных пользователей, атрибуция текстов, авторство сообщений, компьютерная лингвистика, информационная безопасность.

ANONYMOUS WEBSITE USER IDENTIFICATION BASED ON COMBINED FEATURE SET (WRITING STYLE AND TECHNICAL FEATURES)[2]

A. Vorob'yeva[b], A. Gvozdev[b]

[b] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, Alice_w@mail.ru

The task of anonymous web users identification becomes more and more important research task. The number of users is increased dramatically and usage of the Internet for criminal purposes (such as anonymous threats and extremist statements) becomes more frequent. Existing approaches and algorithms for identifying anonymous users are not enough efficient. In the context of this work, user identification means recognizing of an anonymous user on the Internet. Identification is performed by correlating the set of anonymous user features with stored in the database features collected previously. Feature set of the user consists of technical features (IP- address, OS version, etc.) and writing-style features of the user (for short texts in the Russian language). We compared the discriminating power of three feature sets (technical, writing-style and combined) and

[1] Работа выполнена в рамках НИР «Идентификация пользователей порталов сети Интернет».
[2] Done in the framework of S&R work «Identification of Internet portals users»

of three classification methods (Support Vector Machines, neural networks, logistic regression). Results of the experiment showed that the usage of combined feature set (writing-style and technical features) improves the identification accuracy of an anonymous user of the Internet.
Keywords: anonymous users' identification, text attribution, authorship of text messages, author attribution, computational linguistics, information security.

Введение

Вопрос анонимности в Интернете вызывает большое число дискуссий и споров, нельзя сказать однозначно, является ли она положительной или отрицательной его чертой. Необходимость идентификации анонимных пользователей возникает в случае, когда через Интернет были совершены какие-либо противоправные действия, например, в случае, когда производятся анонимные «вбросы» [6] (сознательная дезинформация), либо преступник с помощью Интернет-переписки готовит почву для совершения преступления. Новые методы решения задачи идентификации пользователей могут быть применимы в таких сферах, как компьютерная криминалистика (например, определение автора анонимных злонамеренных сообщений), противодействие терроризму (отнесение текстов с экстремистским содержанием к какому-то известному террористу) и др. [1–13]. Важность задачи идентификации обоснована в работах многих авторов.

В контексте настоящей работы под пользователем портала сети Интернет понимается конкретное физическое лицо, которое своими действиями с ресурсами портала обнаруживает некоторые признаки (характеристики пользователя), определяющиеся особенностями стиля его письменной речи или техническими средствами, которые он использует для доступа в Интернет.

Целью работы является оценка возможности применения различных наборов характеристик и алгоритмов классификации для идентификации пользователей Интернет-порталов.

Задача идентификации пользователя Интернет-порталов

Задача идентификации пользователя в общем случае сводится к решению задачи многоклассовой классификации. Имеется множество объектов – наборов характеристик пользователей $F = \{f1, \ldots, fn\}$, а также множество пользователей $U = \{u1, \ldots, uk\}$, $uj = fk$. Для некоторого подмножества характеристик $F' = \{f_1, \ldots, f_m\} \subseteq F$ известна их принадлежность определенному пользователю, т.е. существует множество пар «характеристики–пользователь». Необходимо установить, кому из множества U действительно принадлежат остальные характеристики $F' = \{f_{m+1}, \ldots, fn\} \subseteq F$, т.е. требуется построить алгоритм a: $F \rightarrow U$, способный классифицировать произвольный набор характеристик из исходного множества – $fi \in F$. Множество U содержит множество «характеристики–известные пользователи», Ftr – обучающая выборка, F' – классифицируемые характеристики.

Характеристики пользователей

При идентификации пользователя одной из главных задач является определение набора характеристик пользователя, выступающих в качестве сопоставляемых ему атрибутов (признаков), и позволяющие получить наиболее высокие показатели качества классификации. Существует достаточно большое количество возможных характеристик, начиная с самых простых технических (IP-адрес и версия операционной системы (ОС) [4]) и лингвистических (частоты слов определенной длины). Также используются более сложные лингвистические характеристики, требующие некоторого синтаксического или семантического анализа [2]. Вторая задача – это разработка методов идентификации пользователя, или классификации набора характеристик анонимного пользователя к известным пользователям. Авторами был отобран ряд моделей для проведения идентификации, в предыдущих исследованиях показавших наилучшую точность классификации [1–13].

Сгруппируем характеристики пользователя $U = \{Ft, F_l\}$ по их типу – технические (F_t), лингвистические (F_l). В табл. 1 приводятся характеристики, используемые в качестве атрибутов (признаков) пользователя, а также описание особенностей их использования при идентификации.

Полагая, что пользователю свойственно использовать несколько устройств для выхода в Интернет, а также сопоставляя ему как конкретному физическому лицу определенные характеристики текста на естественном языке, определим пользователя u_i как кортеж, координатам которого сопоставляется множество характеристик технических средств ft_i и характеристик, обусловленных стилистическими особенностями его письменной речи, fl. Пользователь представляется как

$$u_i = f_i = \{ ft_i, fl_i \},\ u_i \in U,\ f_i \in F,\ ft_i \in Ft, fl_i \in F_l,$$

где f_i – совокупные характеристики пользователя; F – множество возможных совокупностей характеристик пользователей; ft_i – характеристики технических средств пользователя; Ft – множество возможных значений характеристик технических средств; fl_i – лингвистические характеристик текста; F_l – множество возможных значений характеристик текста.

Существуют определенные ограничения при использовании характеристик технических средств пользователя. Для каждого идентифицируемого пользователя они не являются уникальными и характери-

зуют лишь устройство (аппаратное и программное окружение), но никак не самого пользователя. Выбор оптимальной совокупности таких признаков приведен многими авторами, например, в [4, 11].

Отдельно стоит отметить, что при проведении постинцидентного анализа или расследования технические характеристики пользователя получить невозможно, исключение составляют только те случаи, когда сбор этих характеристик ведется на сервере заблаговременно. Лингвистические характеристики этим недостатком не обладают, их использование возможно во всех случаях, пока сообщения пользователя существуют на Интернет-портале и не удалены из базы данных.

Характеристики пользователя	Особенности использования при идентификации
Технические характеристики (F_t)	
Версия ОС Часовой пояс Версия обозревателя Интернет Язык браузера по умолчанию Разрешение экрана	В зависимости от способа получения достоверность определения этих значений может меняться. Легко подделать, поскольку, как правило, их легче всего получить из заголовков HTTP-запроса. Имеет смысл применение активных компонентов на странице для получения более достоверных сведений.
IP-адрес	Извлечение значения этого признака осуществляется на уровне веб-сервера. Несколько физических лиц могут использовать один и тот же IP-адрес 4-й версии. Применение активных компонентов не рассматривается (небезопасно для ОС конечного пользователя).
Лингвистические характеристики (F_l)	
Общее число символов (S) Частота буквенных символов Частота заглавных букв Частота чисел Частота пробелов Частоты управляющих символов Общее число слов (W) Частоты длин слов Частота коротких слов Средняя длина слова Средняя длина предложения в буквах Средняя длина предложения в словах Частота коротких предложений Частота средних предложений Частота длинных предложений Частоты знаков пунктуации Частота использования ссылок Частота использования изображений и другие.	**Выбор характеристик текста в качестве классификационных признаков**. Самыми важными отличительными характеристиками являются слова с самыми высокими частотными характеристиками, т.е. наиболее часто употребляемые автором. **Недостаточное количество и краткость текстов обучающей выборки**. В реальности решение задачи идентификации пользователя усложняется тем, что, как правило, доступно всего лишь несколько достаточно коротких текстов, бесспорно принадлежащих пользователям–кандидатам. **Дисбаланс классов**. Стандартной является ситуация, когда существует неравномерное распределение длин текстов обучающей выборки различных пользователей, что может вызвать некоторый дисбаланс и привести к некорректным результатам. Недостаточная длина текстов обучающей выборки одного пользователя по сравнению с другими не должна снижать вероятность того, что данный пользователь будет выбран как истинный автор текста.

Таблица 1. Характеристики пользователя Интернет-порталов и особенности их использования в задачах идентификации

Эксперимент

Корпус текстов для проведения данного исследования был составлен из текстов на русском языке. Этот корпус был специально составлен для проведения исследований по авторской идентификации пользователя, он представляет собой ограниченную коллекцию текстов различной тематики, для которых автор доподлинно известен, тексты не проходили никакой иной дополнительной обработки. Все тексты относились к различной тематике. Отбор по тематике не производился для того, чтобы эксперимент был максимально приближен к реальной жизненной ситуации, хотя это позволило бы минимизировать влияние тематики на классификацию.

Для составления набора текстов обучающей выборки выбиралось 25 последних сообщений пользователей, все сообщения были произвольной длины. Наиболее частыми были сообщения длиной до 2700 символов (73%), сообщения длиной 2700–5500 и 5500–8200 составляли 10% и 6,5% случаев соответственно. Корпус являлся несбалансированным. Обучающая и тестовая выборки не пересекались.

При выборе технических характеристик для идентификации в комбинации с лингвистическими характеристиками были отобраны только те, которые можно получить этичными методами, безопасными для пользователя, что накладывает определенные ограничения их применения для идентификации исключи-

тельно в неагрессивной среде, поскольку все их значения легко подделать при просмотре веб-страниц. Сопоставление признаков Etag и Cookie с идентификатором пользователя сопряжено с применением специальных методов выработки уникальных значений этих признаков для пользователя, применением специальных средств аутентификации в агрессивной среде, когда пользователь намеренно пытается обойти систему безопасности портала. Такой случай в настоящей работе не рассматривается, как это, например, сделано другими авторами в [4]. Извлечение же MAC-адреса (адреса устройства Media access control) вообще неэтично и сопряжено с техническими трудностями применения активных компонентов на странице; к тому же MAC однозначно определяет только устройство, но не пользователя. При этом несколько лиц могут пользоваться одним и тем же устройством. Для классификации были выбраны следующие признаки, которые легко получить из заголовков HTTP и журнала веб-сервера: версия ОС, версия интернет-браузера, языка интернет-браузера по умолчанию, часовой пояс, разрешение экрана, IP-адрес версии 4.

Для решения задачи идентификации анонимных пользователей было протестировано несколько алгоритмов a: $F \to U$, способных классифицировать произвольный набор характеристик из исходного множества характеристик тестовой выборки: метод опорных векторов (Support Vector Machine, SVM), многослойный перцептрон (NN), наивный байесовский классификатор (NB) и методы логистической регрессии (LR)) [12].

В качестве классификационных признаков использовались характеристики, приведенные в табл. 1. Основой проверки является тестовая выборка, в которой отражено соответствие между документами и их классами. Для проверки качества работы алгоритма классификации необходимо применить его на тестовой выборке и соотнести его решение с заведомо известным правильным решением.

Результаты

При решении задачи идентификации пользователя на качество классификации могут оказывать влияние различные факторы: общее количество пользователей, типы используемых характеристик и сам метод классификации. Именно поэтому далее приводится сравнение точности классификации при использовании различных наборов характеристик в качестве признаков пользователя.

Проверка качества работы алгоритмов классификации производилась на тестовой выборке, в которой проставлено соответствие между характеристиками и пользователями, которым они принадлежат. Численная оценка качества работы алгоритма производилась путем оценки его точности (accuracy). За точность идентификации (A) принимается отношение количества правильно классифицированных наборов характеристик ($f_i \in F'$) – C_c, к общему числу классифицируемых наборов характеристик C_t (размер тестовой выборки) [14].

$$A = \frac{C_c}{C_t} 100\%.$$

Для каждого набора характеристик проводилось тестирование каждого выбранного ранее алгоритма классификации и производилось сравнение с результатами, полученными при использовании других алгоритмов. Результаты для каждого набора характеристик (лингвистические, технические, комбинированные) приведены в табл. 2. Напомним, что во всех случаях используется один тестовый корпус, так что результаты являются сопоставимыми. Использование технических характеристик для идентификации пользователя является традиционным подходом, однако существуют значительные ограничения для использования этих характеристик при решении задачи идентификации пользователя. Идентификация пользователей на основе лингвистических характеристик для русского языка является достаточно новой задачей, требующей дополнительных углубленных исследований, но именно это направление признается наиболее перспективным. Из рисунка и табл. 2 видно, что использование предложенного комбинированного набора характеристик позволяет достичь гораздо большего качества классификации по сравнению с другими характеристиками, точность составляет 90,4%.

	Технические характеристики (*Ft*)	Лингвистические характеристики (*Fl*)	Комбинированные характеристики (*Ft+Fl*)
SVM	72,0	60,0	18,4
NN	72,1	61,6	90,4
NB	66,0	68,8	77,6
LR	68,9	79,2	89,6

Таблица 2. Точность идентификации (%) при использовании различных наборов характеристик пользователя

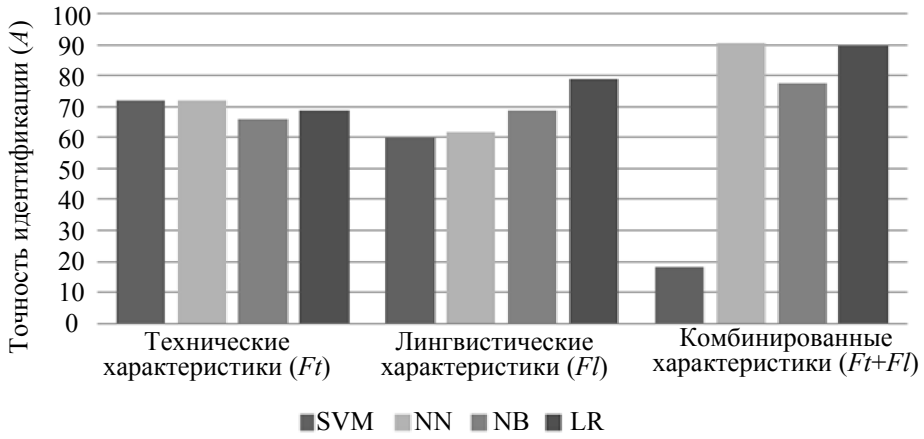

Рисунок. Точность идентификации пользователя при использовании разных наборов характеристик

Заключение

В работе предложен новый подход к решению задачи идентификации автора на основании комбинированного набора характеристик – лингвистических признаков, характеризующих его письменную речь, и признаков, характеризующих терминал его технических средств, которые он использует для доступа к веб-сайту. Предложенный подход может быть использован для классификации в условиях малого объема обучающей выборки. Идентификация пользователей – это один из наиболее характерных примеров, отражающих суть данной проблемы, так как обычно имеется крайне мало сведений о предыдущих характеристиках пользователя, особенно когда дело касается криминалистики. Совместное применение этих признаков позволяет значительно повысить точность идентификации с использованием предложенных алгоритмов. Неожиданно низким является результат применения метода опорных векторов на комбинированных характеристиках. Проведенные эксперименты доказали эффективность идентификации с использованием комбинированного набора характеристик при работе с ограниченной тестовой выборкой.

Сравнение предложенных методов классификации и наборов характеристик показало, что выбор наиболее информативных характеристик является определяющим фактором для получения высокой точности идентификации. С другой стороны, открытым остается вопрос о том, какой из предложенных методов является лучшим. Здесь требуются дальнейшие исследования.

References

1. de Vel A., Anderson O., Corney M., Mohay G. Mining e-mail content for author identification forensics. *SIGMOD Record*, 2001, vol. 30, no. 4, pp. 55–64.
2. Zheng R., Li J., Chen H., Huang Z. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. *Journal of the American Society of Information Science and Technology*, 2006, vol. 57, no. 3, pp. 378–393.
3. Iváncsy R., Juhász S. Analysis of Web User Identification Methods. *International Journal of Computer Science,* 2007, vol. 2, no. 3, pp.172–177.
4. Bessonova E.E, Zikratov I.A., Roskov V.Yu. Analiz sposobov identifikatsii pol'zovatelya v seti Internet [Analysis of Internet User Identification Methods]. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2012, no. 6 (82), pp.128–129.
5. Romanov A.S., Shelupanov A.A., Bondarchuk S.S. Obobshchennaya metodika identifikatsii avtora neizvestnogo teksta [Generalized Authorship Identification Technique]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki,* 2010, no. 1 (21), ch.1, pp. 108-112.
6. Gvozdev A.V., Lebedev I.S., Zikratov I.A. Veroyatnostnaya model' otsenki informatsionnogo vozdeistviya [Probabilistic Analysis Model for Information Influence]. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2012, no. 2 (78), pp. 99–103.
7. Abbasi A., Chen H. Applying Authorship Analysis to Extremist-group Web Forum Messages. *IEEE Intelligent Systems,* 2005, vol. 20, no. 5, pp. 67–75.
8. Park T., Li J., Zhao H., Chau M. Analyzing writing styles of bloggers with different opinions. *Proc. of the 19th Annual Workshop on Information Technologies and Systems (WITS-2009)*. Phoenix, Arizona, USA, 2009, pp. 151–156.
9. Layton R., Watters P., Dazeley R. Authorship attribution for twitter in 140 characters or less. *Second Cybercrime and Trustworthy Computing Workshop (CTC-2010).* Ballart, VIC, Australia, 2010, pp. 1–8.

10. Zheng R., Li J., Chen H., Huang Z. Authorship analysis in cybercrime investigation. *Proc. of the 1st NSF/NIJ conference on Intelligence and security informatics (ISI'03)*. Berlin–Heidelberg, Springer-Verlag, 2003, pp. 59–73.
11. Eckersley P. How Unique is Your Web Browser? *Lecture Notes in Computer Science,* 2010, vol. 6205, pp. 1-18. Available at: https://panopticlick.eff.org/browser-unqiueness.pdf (accessed 26.11.2013).
12. 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.
13. Nawrot M. Automatic Author Attribution for Short Text Documents. *Lecture Notes in Computer Science*, 2011, vol. 6562, pp. 468–477.
14. Manning C.D., Raghavan P., Schütze H. *Introduction to Information Retrieval.* Cambridge University Press, 2008, 504 p. (Russ. ed.: Manning K.D., Ragkhavan P., Shyuttse Kh. *Vvedenie v informatsionnyi poisk.* Moscow, Vil'yams Publ., 2011, 528 p.)

Воробьева Алиса Андреевна – ассистент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, Alice_w@mail.ru

Гвоздев Алексей Вячеславович – ассистент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, a.gvozdev@rcwg.net

Alice Vorob'yeva assistant, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, Alice_w@mail.ru

Aleksei Gvozdev assistant, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, a.gvozdev@rcwg.net