Методика идентификации интернет-пользователя на основе стилистических и лингвистических характеристик коротких электронных сообщений

Technique of web-user identification based on stylistic and linguistic features of short online texts

Воробьева / Vorobeva A.
Алиса Андреевна
(alice_w@mail.ru)
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»,
ассистент.
г. Санкт-Петербург

Ключевые слова: информационная безопасность – information security; идентификация интернет-пользователей – web-user identification; лингвистическая идентификация – forensic linguistics.

В работе рассмотрена задача идентификации пользователей интернет-порталов в условиях возможности деанонимизации источника информации и роста числа преступлений, совершаемых с применением компьютеров и современных средств коммуникации. Представлена методика идентификации интернет-пользователя на основе стилистических и лингвистических характеристик коротких электронных сообщений на русском языке, используемых в качестве идентификатора. Эксперименты показали, что методика позволяет обеспечить более высокую точность идентификации для сообщений на русском языке длиной менее 5000 символов; средняя точность идентификации – 79,7%.

The article touches the problem of web-user identification in the conditions of possibility to hide the identity (the source of the information) and increasing number of crimes committed with the use of computers and mobile devices. In this paper is proposed the technique of web-user identification based on stylistic and linguistic features of online texts in Russian. Experiments show that proposed technique provides higher identification accuracy for the short online messages in Russian (length less than 5000 symbols); average accuracy achieved - 79.7%.

Введение

Идентификация и аутентификация субъектов информационных процессов – это одна из базовых задач обеспечения информационной безопасности. В работе рассматриваются информационные процессы, протекающие в такой среде как Интернет и реализуемые с помощью коммуникационных средств интернет-ресурсов (веб-сайтов, социальных сетей, форумов). При информационном обмене в Интернет крайне важным является возможность идентифицировать или аутентифицировать пользователя, определить является ли пользователь тем, за кого он себя выдает. Важность задачи идентификации пользователей в Интернет обоснована в работах многих авторов [1–3].

Одним из наиболее перспективных направлений развития технологий идентификации является биометрическая идентификация. К нему относятся и методы лингвистической идентификации, основанные на лингвистических или стилистических характеристиках электронных сообщений. Данные методы могут применяться самостоятельно или как дополнительный скрытый этап при идентификации для получения доступа к ресурсам и сервисам публикации электронных сообщений на веб-ресурсе. Лингвистическая идентификация также позволяет осуществлять противодействие использованию интернет-ресурсов криминальными и террористическими организациями для анонимных угроз, распространения экстремистских или террористических идеологий, распространения материалов противоправного содержания [4, 5].

Предыдущие исследования по лингвистической идентификации относятся к двум группам: это методы идентификации автора литературных произведений и методы идентификации пользователей по коротким электронным сообщениям на иностранных языках. Существует ряд ограничений, которые не позволяют применить их для идентификации по коротким электронным сообщениям: это длина и язык текстов, для которых данные методы разрабатывались. Максимальная точность идентификации по коротким сообщениям на русском языке длиной 5000 символов составляет 47% [2], в работе [5] для сообщений длиной 500 слов получена точность 55%.

Под точностью идентификации (A) понимается отношение количества правильно идентифицированных пользователей для сообщений – $IdentU_{corr}$ к общему числу сообщений тестовой выборки – $|T_{tr}|$.

$$A = \frac{IdentU_{corr}}{|Ttr|}$$

Существующие методы лингвистической идентификации интернет-пользователей обладают рядом

недостатков и ограничений, точность идентификации является достаточно низкой. Возникает задача повышения точности идентификации и аутентификации интернет-пользователей.

Задача идентификации интернет-пользователя

Имеется некоторое поступившее от пользователя u_k сообщение t_j и набор потенциальных пользователей $U_{потенц}$. Для каждого потенциального пользователя необходимо определить вероятность, с которой он является автором поступившего сообщения, т.е. необходимо классифицировать данное сообщение к имеющимся пользователям.

Каждое сообщение представлено как набор характеристик $t_j = F_j = (f_{j1},...,f_{jn})$, где n – число характеристик или идентификационных признаков. Пользователь u_k представляется, как множество его сообщений $u_k = T_k = (t_{k1},...,t_{kl})$ где l – количество сообщений пользователя u_k.

Задача идентификации интернет-пользователей делится на две подзадачи:

1. Сбор и формирование базы характеристик потенциальных пользователей (базы данных эталонных моделей представления пользователя), содержащей эталонные шаблоны пользователей U (КММППэталоны), в том числе потенциальных пользователей $U_{потенц}$.

2. Идентификация и аутентификация пользователя по поступившему новому электронному сообщению t_j (производится разбор сообщения и сравнение с эталонными шаблонами пользователей). В случае успешного сравнения личность пользователя и ее подлинность можно считать установленными.

Идентификация производится путем соотнесения набора выявленных характеристик электронного сообщения с характеристиками сообщений, собранными ранее и уже имеющимися в базе данных, для которых известна их принадлежность определенному пользователю. Имеется множество сообщений различных пользователей: $T = \{t_1,...,t_z\}$, где z – количество сообщений. Множество пользователей $U = \{u_1,...,u_y\}$, где y – количество пользователей. Для $T' \in T$ сообщений известна их принадлежность определенному пользователю, т.е. существует некоторое множество, состоящее из пар соответствий пользователей и сообщений. Каждый пользователь представляется, как $u = T = (i_1,...,i_l)$, где l – количество сообщений пользователя u.

Требуется построить алгоритм $a: t_j \to U_{потенц}$, способный определить с какими вероятностями сообщение t_j принадлежит каждому пользователю из $U_{потенц}$, где t_j – сообщение по характеристикам которого необходимо произвести идентификацию (предъявленный идентификатор). Необходимо произвести обучение данного алгоритма по обучающей выборке $T_{tr} \in T$ и его верификацию по $T_{test} \in T$, где T_{tr} – обучающая выборка, T_{test} – тестовая выборка. Далее обученный алгоритм A должен быть способен классифицировать поступившее сообщение t_j к пользователям $U_{потенц}$ Искомым пользователем является пользователь, для которого P (t_j принадлежит u_{max}) является максимальной.

Методика идентификации интернет-пользователя на основе стилистических и лингвистических характеристик коротких электронных сообщений

В качестве идентификационных признаков в работе используются как частотные (частоты использования определенных символов, знаков препинания, изображений, гиперссылок, частоты слов определенной длины, частоты служебных слов и пр.), так и статистические характеристики текста (длина сообщения, средняя длина слова, средняя длина предложения, час и день публикации сообщений, и пр.). Одному пользователю в некоторый промежуток времени свойственны определенные характерные особенности. Важным является тот факт, что данные характеристики должны быть внесены в текст подсознательно (например, пол лица, от которого ведется повествование, не может выступать в качестве признака пользователя). Множество идентификационных признаков составляет комплексную многоуровневую модель представления пользователя (КММПП). В отличие от предыдущих работ, предлагается использовать максимальное количество идентификационных признаков пользователя, экспериментально было доказано, что для разных частных задач идентификации наиболее информативные признаки различаются. Подробное описание КММПП и используемых признаков приведено в [3]. КММПП отличается от известных использованием ряда уникальных характеристик электронных текстовых сообщений и позволяет производить идентификацию пользователей по характеристикам электронных текстовых сообщений длиной менее 5000 символов, эксперименты подтвердили более высокую точность идентификации с использованием КММПП [11].

На основе разработанной КММПП предложена методика идентификации интернет-пользователя. Методика позволяет осуществлять идентификацию и аутентификацию интернет-пользователей, может быть применима в качестве части системы разграничения доступа (СРД). Применение разработанной методики в СРД к сервисам публикации электронных сообщений позволяет дополнить стандартные механизмы идентификации/аутентификации процедурой дополнительной скрытой идентификации. Таким образом, скрытая идентификация является дополнительным средством защиты для обеспечения информационной безопасности информационных процессов, протекающих в Интернет.

Методика идентификация включает в себя два основных этапа (рис. 1):

1. Формирование КММПП и ДСПП для u_k и $U_{потенц}$.

1.1. Извлечение характеристик, входящих в КММПП, из сообщения t_j, и формирование КММППu_k пользователя, размещающего сообщение t_j.

1.2. Извлечение КММПП$_{потенц}$ из БД эталонных КММПП.

1.3. Формирование ДСППu_k и ДСПП$_{потенц}$ путем динамического вычисления и отбора наиболее информативных признаков из КММПП$_{потенц}$ и КММППu_k, по разработанному методу формирования ДСПП [7].

2. Идентификацию пользователя, по электронному сообщению t_j:

2.1. Дискретизация непрерывных признаков из ДСПП$_{потенц}$ и ДСППuu_k, формирование ДСПП'u_k и эталонных ДСПП'$_{потенц}$.

2.2. Сравнение ДСПП'u_k с эталонными ДСПП'$_{потенц}$.

2.3. Принятие итогового решения о подлинности предъявленного идентификатора.

Первый и второй шаги второго этапа производятся по разработанному методу сравнения ДСПП'u_k с эталонными ДСПП'$_{потенц}$ потенциальных пользователей на основе метода Случайного Леса (*Random Forest*) [6], включающему предварительную дискретизацию идентификационных признаков из ДСПП$_{потенц}$ и ДСППu_k.

На третьем шаге методики определяется пользователь, которому с наибольшей вероятностью принадлежит поступившее электронное сообщение, из числа потенциальных пользователей путем сравнения ДСПП'u_k с ДСПП'$_{потенц}$.

Определяется идентификатор пользователя с наибольшей вероятностью, являющегося автором сообщения P (t_j принадлежит $u_{\max}$). И производится сравнение предъявленного идентификатора с определенным идентификатором, если идентификаторы совпали, то предоставляется доступ к ресурсу, т.е. к размещению сообщения. Иначе делается отметка о попытке НСД, далее администратор системы обрабатывает данные отметки.

Стоит отметить, что идентификация Интернет-пользователей при информационном обмене электронными сообщениями может выполняться в виде верификации, аутентификации или распознавания. При верификации подтверждается идентичность ДСППu_k, сформированного по сообщению t_j, и эталонного шаблона пользователя u_k, хранимого в базе данных. Аутентификация – подтверждает соответствие ДСППu_k, сформированного по сообщению t_j, одному из шаблонов, хранящихся в базе данных. При распознавании, если полученный ДСППu_k и один из хранимых шаблонов оказываются одинаковыми, то система идентифицирует пользователя с соответствующим шаблоном.

Эксперименты и оценка результатов

Сравнительный анализ итоговой точности идентификации с использованием разработанной методики проводился на корпусе сообщений, составленном из текстов, входящих в Национальный корпус русского языка, и текстов записей блогов на русском языке (портала LiveJournal). Корпус представляет собой коллекцию текстов, для которых пользователь, являющийся автором сообщения доподлинно известен. Данный корпус ранее использовался в работах [3], [6], [7]. Были отобраны 3700 сообщений 166 пользователей с максимальной длиной 4986 символов. Большинство сообщений были длинной до 498 символов. Составлены наборы тестовых данных с различным уровнем несбалансированности и количества сообщений. Соотношение минимального и максимальных

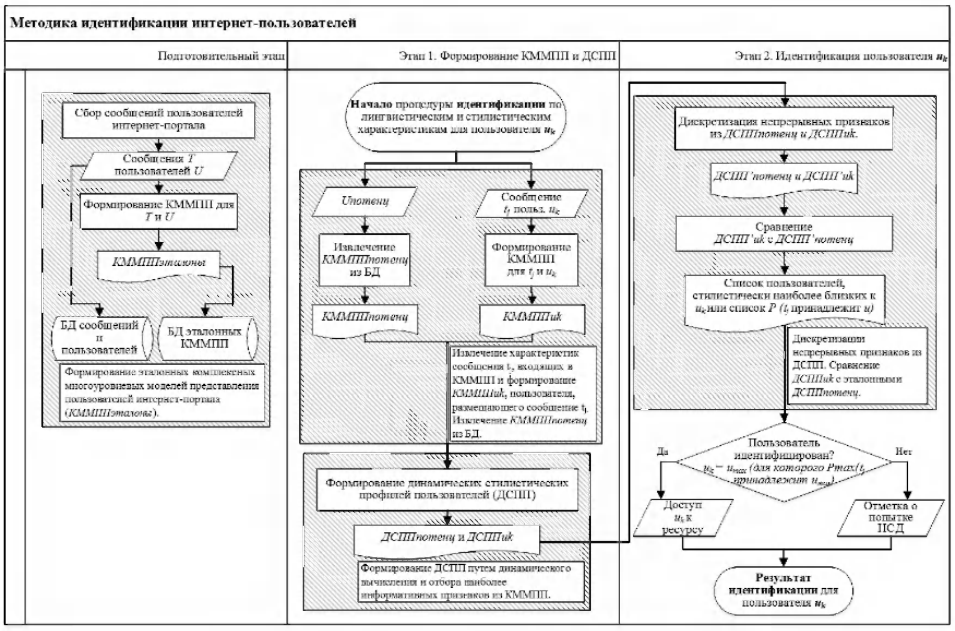

Рис. 1. Блок-схема методики идентификации интернет-пользователя

Таблица 1

Точности идентификации с использованием разработанной методики

Количество сообщений одного пользователя (мин:макс)	Точность идентификации (*A*), %
24:25	80,85
20:25	80,63
10:20	80,9
5:25	79,68
10:10	78,18
8:10	78,19
5:10	78,71
2:10	80,44

ного количества обучающих примеров на одного пользователя приведено в Таблице 1 (мин:макс).

Проведенные эксперименты показали, что условиях ограниченной длины текстов, их малого количества и неравномерного распределения по пользователям разработанная методика обеспечивает боле высокую точность идентификации. Точность идентификации по базовой КММПП на малом количестве текстов несколько ниже точности на нормальном количестве, то при идентификации по разработанной методике точность является одинаково высокой (Таблица 1).

Предложенная методика позволяет производить идентификацию интернет-пользователя (в условиях ограниченной длины текстов и несбалансированности обучающей выборки) со средней точностью на малом количестве текстов – 78,88%, на нормальном количестве текстов – 80,52% по электронным текстовым сообщениям длиной до 5000 символов. Итоговая средняя точность 79,70% примерно на 32,70% выше, чем существующие статистические методы [2].

Заключение

В работе предложена методика идентификации интернет-пользователя, которая может быть использована для повышения безопасности информационных процессов Интернет. Данная методика позволяет повысить точность идентификации, что дает возможность использовать результаты в СРД к сервисам публикации электронных сообщений, как часть систем предотвращения утечек информации при расследовании инцидентов информационной безопасности, а также для выявления потенциальных злоумышленников Интернет до совершения ими противоправных действий (угрозы, шантаж, призывы к экстремизму, терроризму и к межнациональной розни, сознательная дезинформация). Анализ точности, достигаемой в условиях ограниченной длины текстов и несбалансированности обучающей выборки и точности, достигнутой в работах других исследователей, показал, что предложенная методика позволяет производить идентификацию интернет-пользователя со средней точностью на малом количестве текстов – 78,88%, на нормальном количестве текстов – 80,52% по электронным текстовым сообщениям длиной до 5000 символов. Итоговая средняя точность 79,70% примерно на 32,70% выше, чем существующие статистические методы.

Литература

1. Лебедев, И. С.Методика идентификации авторства текстов коротких сообщений пользователей порталов сети интернет на основе методов математической лингвистики / И.С. Лебедев, М.Е. Сухопаров // В мире научных открытий. – 2014. – Т. 6.1. – № 54. – С. 599–622.

2. Романов, А. С. Методика и программный комплекс для идентификации автора неизвестного текста: автореф. дисс. ... канд. тех. наук : 05.13.18 / А. С. Романов. – Томск, 2010.– 26 с.

3. Vorobeva, A. Forensic linguistics: automatic web author identification / A. Voroveba // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – Т. 16, № 2(102). – С. 295–302.

4. Authorship attribution for social media forensics / A. Rocha [et al.] // IEEE Transactions on Information Forensics and Security.– 2017.– Vol. 12, No. 1. – P. 5–33.

5. Doppelganger finder: Taking stylometry to the underground / S. Afroz [et al.] // 2014 IEEE Symposium on Security and Privacy (SP). – 2014. – P. 212–226.

6. Vorobeva, A. Examining the Performance of Classification Algorithms for Imbalanced Data Sets in Web Author Identification / A. Voroveba // Proceedings of the 18th Conference of Open Innovations Association FRUCT. – 2016. – P. 385–390.

7. Воробьева, А. А. Отбор информативных признаков для идентификации интернет-пользователей по коротким электронным сообщениям / А.А. Воробьева // Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Т. 17, № 1 (107). – С. 117–128.