УДК 004.931

СПОСОБ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ В СЕТИ ИНТЕРНЕТ

Е.Е. Бессонова, И.А. Зикратов, Ю.Л. Колесников, В.Ю. Росков

Рассматриваются механизмы идентификации пользователей в сети Интернет. Предложен способ формирования признакового пространства для идентификации пользователя, обоснован метод идентификации по вторичным характеристикам рабочей среды. Для проверки полученных результатов проведен вычислительный эксперимент.
Ключевые слова: идентификация, информативность, признак, кортеж, пользователь.

Введение

Одной из важных задач в теории защиты информации является задача идентификации пользователя в сети Интернет. Актуальность данной задачи обусловлена целесообразностью идентификации субъектов сети при построении системы защиты информации, в частности, для выявления нарушителей.

Целью данной работы является определение рационального признакового пространства и способа идентификации, позволяющего повысить достоверность отождествления пользователей с имеющимися записями в базе данных информационного ресурса.

Для современных информационных систем применяются способы идентификации, основанные на хранении IP-адресов компьютеров посетителей и записи на компьютер пользователя данных Cookie. К недостаткам первого способа относится широкая распространенность динамических IP-адресов, выделяемых из пула провайдера в момент подключения пользователя, а также возможность использования в сети прокси-серверов, анонимайзеров и механизма NAT (Network Address Translation), что снижает степень достоверности идентификации пользователя [1]. Недостатком второго способа является привязка Cookie к конкретному браузеру, что снижает достоверность идентификации при использовании нескольких браузеров. Другим недостатком использования данной технологии является возможность подмены и уничтожения данных Cookie, а также отключения самого механизма пользователем.

Таким образом, оба способа не позволяют в ряде случаев достичь требуемой степени достоверности идентификации [2]. В то же время существуют способы получения данных, характеризующих рабочую среду пользователя. Под рабочей средой пользователя понимаются данные об операционной системе пользователя, шрифтах, параметрах экрана, плагинах, посещенных ссылках и т.п. Известны попытки использования перечисленных данных в качестве признаков идентификации [3]. Однако использование такой технологии влечет за собой увеличение объема трафика, что приводит к возрастанию времени загрузки сайта.

Таким образом, задача состоит в разработке способа, позволяющего осуществить рациональный выбор признаков, необходимых для повышения степени достоверности идентификации пользователя в сети Интернет.

Обоснование рационального признакового пространства

Сформулированная задача решена в два этапа:

1. сбор и обработка данных с целью обоснования рационального признакового пространства;
2. обоснование метода идентификации пользователя в выбранном пространстве признаков.

В рамках первого этапа при помощи тестового сайта был произведен сбор данных о рабочей среде пользователя. Для этого при каждом посещении пользователя сайта собирались следующие данные: время посещения; настоящий (контрольный) идентификатор пользователя (пользователь вводил логин и пароль); IP-адрес пользователя; строка-идентификатор User Agent; набор плагинов браузера, предоставляемый при помощи технологии Javascript, а также предоставляемая при помощи технологии Javascript информация о браузере и операционной системе, о языке операционной системы и разрешении экрана; список установленных шрифтов, собранный при помощи технологий ActiveX и Flash [3].

Все собранные признаки можно разделить на программные и аппаратные. К аппаратным можно отнести MAC-адрес, получаемый при помощи технологии Java, к программным – все остальные.

Совокупность перечисленных признаков (идентификаторов), кроме контрольных идентификаторов, вводимых пользователем, получила название кортежа. Именованный кортеж признаков, относящийся к конкретному пользователю, в данной работе называется профилем пользователя.

Очевидно, что тот или иной признак в различной мере способствует процессу отождествления кортежа с тем или иным профилем. Для выявления наиболее значимых для идентификации признаков введено понятие информативности признака. Под информативностью в работе понимается степень влияния признака в кортеже идентифицируемого объекта на результат отождествления с имеющимися профилями пользователей.

Для вычисления меры информативности каждого признака в работе использован метод регрессионного анализа. В качестве зависимой переменной было выбрано расстояние между кортежами признаков, в качестве набора независимых – набор бинарных расстояний между значениями признаков в

отдельности. Полученные в результате регрессии коэффициенты при признаковых расстояниях были приняты за коэффициенты информативности для этих признаков [4].

Технология	Признак	Информативность
ETag(кэш браузера)	Идентификатор	0,88765812
Supercookie	Идентификатор	0,758318026
Cookie	Идентификатор	0,692106732
Java	MAC	0,507266254
IP	IP	0,50545622
Flash	Шрифты	0,36639634
Javascript	Плагины	0,320531032
ActiveX	Шрифты	0,310195763
CSS	Параметры экрана	0,230771305
TCP-протокол	Операционная система	0,185970125
Браузерные особенности	Браузер	0,118123798
Java	Шрифты	0,117587794
Javascript	Браузер	0,081847248
Javascript	Параметры экрана	0,069818299
Javascript	Часовой пояс	0,06374125
Javascript	Язык	0,035509264
User Agent	Браузер	0,029075342
User Agent	Операционная система	0,028753274
Java	Операционная система	0,0214338177
User Agent	Язык, кодировка	0,009661464
CSS + History + JS	Посещенные ссылки	0,006501557
Javascript	Операционная система	0,000394953

Таблица. Значения информативности признаков кортежа

Используемое в качестве критерия идентификации пользователей регрессионное уравнение для псевдорасстояния между двумя кортежами признаков имеет вид

$$Score(U_1, U_2) = \sum D_i \cdot a_i ,$$

где U_1 и U_2 – кортежи признаков; D_i – бинарное расстояние между *i*-ми признаками из кортежей; a_i – коэффициент информативности *i*-го признака; $Score$ – профиль с минимальным значением, принимающийся за соответствующий эталону профиль. При этом установлен эмпирический порог $Score_{\max} = 3,47$, при превышении которого эталон считается не подходящим ни к одному профилю. В этом случае создается новый профиль. Под бинарным расстоянием между кортежами признаков, а также между значениями одного и того же признака понимается величина, принимаемая за 0 при совпадении кортежей и за 1 – при несовпадении. Полученные в результате регрессионного анализа значения информативности приведены в таблице.

Вывод о достоверности расчетов сделан на основе величины фактического F-критерия Фишера, а также величин P-значений. Величина фактического F-критерия Фишера оказалась близка к нулю, из чего сделан вывод о статистической значимости полученных в результате регрессионного анализа коэффициентов. Полученные величины P-значений также малы, из чего следует, что вероятность сделать ложный вывод на основе регрессионного уравнения близка к нулю.

Обоснование метода идентификации пользователя в пространстве признаков

На втором этапе рассматривались методы идентификации пользователя по его кортежу путем сравнения его с накопленной базой.

Для решения задачи рассматривалось несколько методов:

- байесовский классификатор для вычисления вероятности идентичности профиля с эталоном;
- метод расчета корреляции между кортежами признаков;
- метод расстояния Левенштейна;
- метод прямого сравнения с эталоном.

Сравнение методов осуществлялось по двум параметрам: скорость и надежность. Надежность оценивалась как процент верно идентифицированных пользователей при увеличении процента измененных данных. За скорость принималось время работы алгоритма в секундах. Сравнительный результат работы методов представлен на рис. 1, 2.

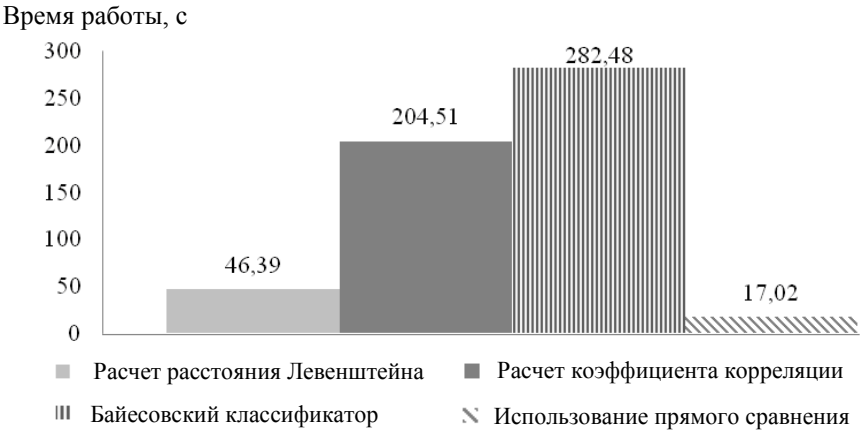

Рис. 1. Сравнение скорости алгоритмов идентификации:

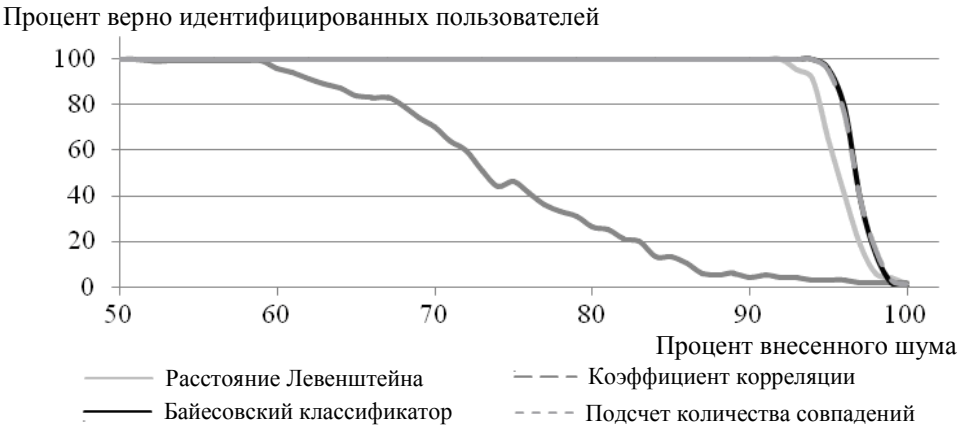

Рис. 2. Сравнение надежности алгоритмов идентификации. Искажения 50–100%

Как видно из графиков, метод прямого сравнения с эталоном показал результаты, по надежности сравнимые с байесовским классификатором, имея при этом самую высокую скорость работы из представленных методов. Кроме этого, он подходит для расчета бинарных расстояний между признаками. Именно он и был использован при идентификации пользователей.

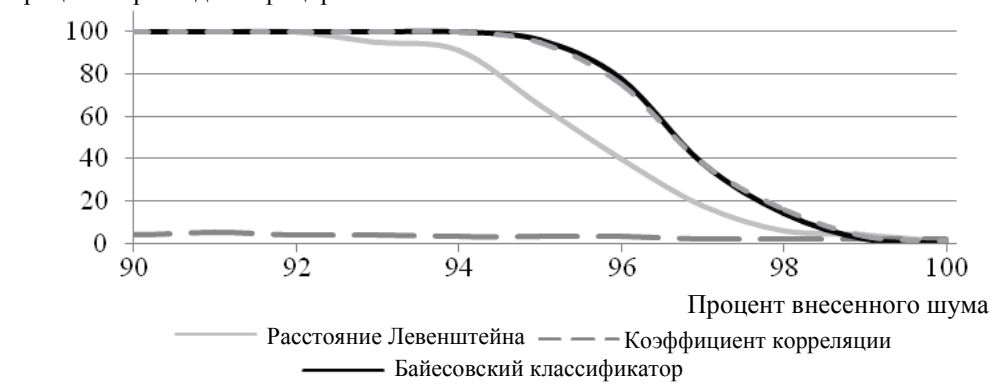

Рис. 3. Сравнение надежности алгоритмов идентификации. Искажения 90–100%

Проведение эксперимента

Для проверки полученных показателей информативности был проведен вычислительный эксперимент с целью оценки степени достоверности идентификации по кортежу признаков. В качестве входных данных были использованы учетные записи, выбранные в случайном порядке (эталоны); статистика учетных записей пользователей, заходящих на тестовый сайт не менее двух раз; признаки, упорядоченные по возрастанию информативности.

Цель эксперимента – определить зависимость степени достоверности идентификации от количества признаков, включенных в профиль пользователя. Для эксперимента был взят полный кортеж признаков, описанный в таблице. При помощи этого кортежа проводилось сравнение эталонов с пользовательскими профилями, полученными в результате сбора статистики. После этого суммарная информативность уменьшалась за счет удаления из кортежа признаков с рассчитанной наибольшей информативностью, затем с наименьшей информативностью. Результаты эксперимента отображены на графиках (рис. 4).

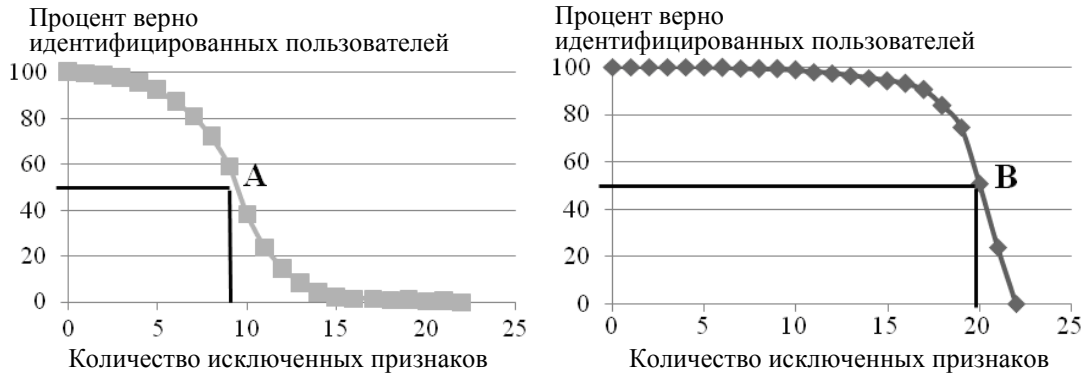

Рис. 4. Графики зависимостей достоверности идентификации от количества исключенных из кортежа признаков: сначала с высокой информативностью (а); сначала с низкой информативностью (б). Точки А и Б – точки, в которых происходит падение точности идентификации до уровня 50%

По горизонтальной оси располагается количество признаков, включенных в профиль пользователя, по вертикальной оси – процент правильно идентифицированных пользователей.

Оба графика монотонно убывают, что объясняется тем, что с уменьшением информативности уменьшается и процент распознанных эталонов. Прямой линией зафиксирован уровень, при котором происходит снижение точности идентификации в 2 раза. В точке А (рис. 4, а), находящейся на пересечении прямой и графика, видно, что зафиксированный уровень точности идентификации сохраняется при удалении до 9–10 самых информативных признаков, что подтверждает гипотезу об информативности признаков.

Из результата эксперимента следует, что наиболее информативными следует считать признаки, полученные с помощью технологий ETag (информативность равна 0,888), Supercookie (информативность равна 0,758). Признаки с меньшей информативностью – Cookie (информативность равна 0,692), MAC-адрес (информативность равна 0,507) и IP-адрес (информативность равна 0,505) также обладают высокой степенью информативности, однако их подмена является менее затратной для злоумышленника. Как видно из графика, одновременное удаление из кортежа вышеперечисленных признаков резко снижает степень достоверности идентификации.

На втором графике (рис. 4, б) также зафиксирован уровень, при котором происходит снижение точности идентификации в 2 раза. Точка В, находящаяся на пересечении прямой и графика, расположена дальше от начала координат, что обусловливается удалением из кортежа максимально информативных признаков в последнюю очередь.

Результаты расчета подтверждают, что степень достоверности идентификации зависит от набора признаков в кортеже. Был использован доверительный уровень степени идентификации в 95%. При отбрасывании, в первую очередь, наименее информативных признаков пересечение уровня в 95% происходит при переходе границы в 8 признаков. Отсюда следует, что рациональным признаковым пространством для идентификации является наличие в кортеже восьми наиболее информативных идентификаторов: ETag, Supercookie, Cookie, MAC, IP, шрифты через Flash, плагины, шрифты через ActiveX. По сравнению с Cookie данный кортеж обеспечивает в 6,3 раза большую информативность (4,35 против 0,69).

Большая часть указанных в работе признаков применима для идентификации пользователей на мобильных устройствах. Очевидно, что при использовании мобильных устройств, планшетов, виртуальных машин и т.п., возможно изменение признакового пространства, однако сам подход остается прежним. Также предложенный механизм позволяет не противоречить системам идентификации на основе OpenID и социальных профилей.

Таким образом, задача повышения степени достоверности идентификации по сравнению с используемыми в настоящее время механизмами решена.

Заключение

Для нейтральной среды (т.е. для пользователей, не стремящихся фальсифицировать идентификацию) будет целесообразно использовать только наиболее информативные признаки, не требующие подтверждения пользователя и дополнительных запросов к серверу.

Результаты исследований могут быть использованы для автоматизированной оптимизации систем обнаружения вторжений при выставлении адаптивного порога проверки для обнаружения объекта, который был ассоциирован с нарушителем.

Если среда использования является агрессивной или затруднено получение наиболее информативных признаков, то представляется возможным использование всего кортежа, либо набора информативных признаков. Целью дальнейшей работы авторов является исследование по определению показателей качества идентификации по вторичным признакам при использовании нарушителем наиболее распространенных способов маскировки – подмены и (или) удаления признаков.

Литература

1. Understanding IP Addressing: Everything You Ever Wanted To Know [Электронный ресурс]. – Режим доступа: http://web.archive.org/web/20100821112028/http://www.3com.com/other/pdfs/infra/corpinfo/en_US/501302.pdf, свободный. Яз. англ. (дата обращения 15.10.2011).
2. McKinkley K.: Cleaning Up After Cookies. iSec Partners White Paper [Электронный ресурс]. – Режим доступа: http://www.isecpartners.com/storage/white-papers/iSEC_Cleaning_Up_After_Cookies.pdf, свободный. Яз. англ. (дата обращения 15.10.2011).
3. Кантор И. Способы идентификации в интернете [Электронный ресурс]. – Режим доступа: http://javascript.ru/unsorted/id, свободный. Яз. рус. (дата обращения 15.10.2011).
4. Таха Х.А. Введение в исследование операций. – 2-е изд. Пер. с англ. – М.: Вильямс, 2005. – 912 с.

Бессонова Екатерина Евгеньевна – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, аспирант, bessonova@cit.ifmo.ru

Зикратов Игорь Алексеевич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, доктор технических наук, профессор, зав. кафедрой, zikratov@cit.ifmo.ru

Колесников Юрий Леонидович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, доктор физ.-мат. наук, профессор, проректор, kolesnikov@mail.ifmo.ru

Росков Владислав Юрьевич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, студент, vos@vos.uz