10 МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ

УДК 004.931

РАЗЛИЧЕНИЕ ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ ИХ ПОВЕДЕНИЯ В СЕТИ ИНТЕРНЕТ

Д.С. Юрасов, И.А. Зикратов

Работа посвящена идентификации и аутентификации пользователей в web-пространстве. Предложен способ различения двух пользователей, осуществляющих доступ в сеть Интернет с одного общего компьютера и браузера, основанный на анализе истории посещений сайтов каждым из пользователей. Способ базируется на выявлении часто и регулярно посещаемых сайтов и их последующей кластеризации иерархическим методом. Для проверки работоспособности способа был проведен анализ истории посещения сайтов более чем 260 000 пользователей, собранной с помощью cookie-идентификаторов в сети Интернет, которые хранятся на их компьютерах. Эксперимент показал высокий процент верно идентифицированных пользователей.

Ключевые слова: защита информации, иерархическая кластеризация, cookies, многопользовательский компьютер, идентификация пользователей, аутентификация.

Введение

Одной из важных задач в теории защиты информации является задача идентификации пользователя в сети Интернет. Идентификация и аутентификация являются важнейшими задачами, решаемыми при построении систем разграничения доступа, которые, в свою очередь, играют ключевую роль в системах обеспечения информационной безопасности.

В современных информационных системах, как правило, применяются способы идентификации, основанные на информации о IP-адресах компьютеров посетителей и технологии Cookie [1, 2]. Cookies представляют собой данные небольшого объема, которые создаются при посещении пользователями веб-сайтов, хранятся на их компьютерах и содержат информацию, например, о настройках веб-сайта или о профиле. Усовершенствованиями этих способов являются решения, предложенные в работах [3–5], которые заключаются в анализе служебной информации о компьютере пользователя и выборе наиболее значимых признаков. Они позволили существенно увеличить степень достоверности идентификации пользователей при отсутствии информации о IP-адресах или cookies. Однако всем этим способам присущ существенный недостаток: они отождествляют пользователя и его компьютер или браузер, что не всегда корректно. По результатам последних исследований, доля многопользовательских компьютеров (т.е. таких, за которым работают два и более пользователей) варьируется от 20 до 40% [6].

Целью настоящей работы является разработка способа, позволяющего различить пользователей, осуществляющих доступ в сеть Интернет посредством одного компьютера и браузера. Высокий процент многопользовательских компьютеров говорит об актуальности этой задачи, однако, среди открытых источников печати публикации, в которых предложен метод ее решения, не представлены.

Предложенный способ заключается в выявлении часто и регулярно посещаемых пользователем сайтов и их последующей кластеризации. Для решения поставленной задачи в работе использованы данные о посещении пользователем сайтов в сети Интернет, которые могут быть либо собраны с помощью Cookie или других технологий, либо получены из сторонних источников.

Способ различения пользователей с общими компьютером и браузером

Подготовка данных. Для сбора информации о посещении пользователями сайтов использовались данные Cookie, поэтому далее отождествляем пользователей с браузером.

Как правило, системы разграничения доступа не располагают информацией о количестве людей, пользующихся конкретными компьютером и браузером. Чтобы решить эту проблему при подготовке данных, объединим истории посещений сайтов для случайно выбранных пар cookies. Всего было объединено более 130 000 пар пользователей. Такие пары называем «склеенными», как и истории посещений сайтов, полученные объединением историй посещений исходных пользователей. В результате были получены данные, в которых за каждым cookie-идентификатором скрывается минимум два пользователя. Такие cookies называем многопользовательскими.

Значимые сайты. В основу идеи различения пользователей положен тот факт, что у большинства пользователей существуют сайты, которые они посещают достаточно часто и регулярно. Для выявления таких сайтов история посещений каждого пользователя была разбита на сессии. Под пользовательской сессией понимается упорядоченная по времени последовательность посещений сайтов, в которой временной интервал между соседними посещениями сайтов не превышает 30 минут. Сессии объединенных пар cookies были пронумерованы случайным образом, но так, чтобы исходный хронологический порядок

сессий «несклеенных» пользователей сохранился. Будем говорить, что сайт является значимым для пользователя, если выполнены два условия:

1. он встретился хотя бы в десятой части всех известных сессий;
2. если пронумеровать в хронологическом порядке все известные сессии пользователя, то дисперсия номеров сессий, в которых встретился данный сайт, больше определенного числа.

Первое условие гарантирует частоту посещения сайта, а второе – регулярность, исключая сайты, к которым пользователь проявил лишь локальный интерес, т.е. интенсивно интересовался ими лишь на протяжении некоторого относительно короткого интервала времени.

В изучаемых искусственных данных у всех пользователей нашелся хотя бы один «значимый» сайт. Более 88% обычных, «несклеенных» пользователей имеют хотя бы один «значимый» сайт. Для всех сайтов было определено количество пользователей, для которых данный сайт является «значимым». Самыми популярными такими сайтами оказались vk.com, e.mail.ru и odnoklassniki.ru.

Кластеризация значимых сайтов. Экспериментальные данные были подготовлены таким образом, что за каждым cookie-идентификатором в них скрываются как минимум два реальных пользователя, следовательно, и значимые сайты должны для них делиться минимум на две соответствующие группы; далее предполагаем, что их ровно две.

Рассмотрим конкретную cookie *u*. Пусть у нее было всего *n* сессий, ее значимыми сайтами являются $s_1,...,s_m$. Для всех *j* от 1 до *m* и всех *k* от 1 до *n* сайту s_j поставим в соответствие вектор $s_{j1},...,s_{jn}$, где s_{jk} принимает значение 1, если пользователь *u* посещал сайт s_j в *k*-ю сессию, и 0 иначе.

В качестве расстояния между «значимыми» сайтами s_i и s_j пользователя *u* использовалось расстояние Жаккара *J*, которое является одной из самых распространенных бинарных мер сходства объектов. Оно вычисляется по формуле

$$J = \frac{M_i + M_j}{M_i + M_j + M_{ij}},$$

где M_{ij} – количество сессий пользователя *u*, в которых присутствуют и s_i, и s_j; M_i – количество сессий пользователя *u*, в которых присутствует сайт s_i, но нет сайта s_j; M_j – количество сессий пользователя *u*, в которых присутствует сайт s_j, но нет сайта s_i [7, 8]. Из определения расстояния Жаккара следует, что в идеальной ситуации расстояние между значимыми сайтами разных пользователей «склеенной» cookie *u* будет близко к 1, а между сайтами одного пользователя – к 0, что является еще одной причиной для выбора именно этой меры сходства объектов.

Далее сайты первой категории cookie *u* были разделены на две группы с помощью иерархической кластеризации, с использованием введенной меры сходства. В качестве расстояния между кластерами было выбрано арифметическое среднее попарных расстояний между сайтами, так как оно является менее чувствительным, по сравнению с другими общепринятыми методами расчета, к выбросам и искажениям в данных, которые, учитывая специфику сбора информации, не исключены [9, 10].

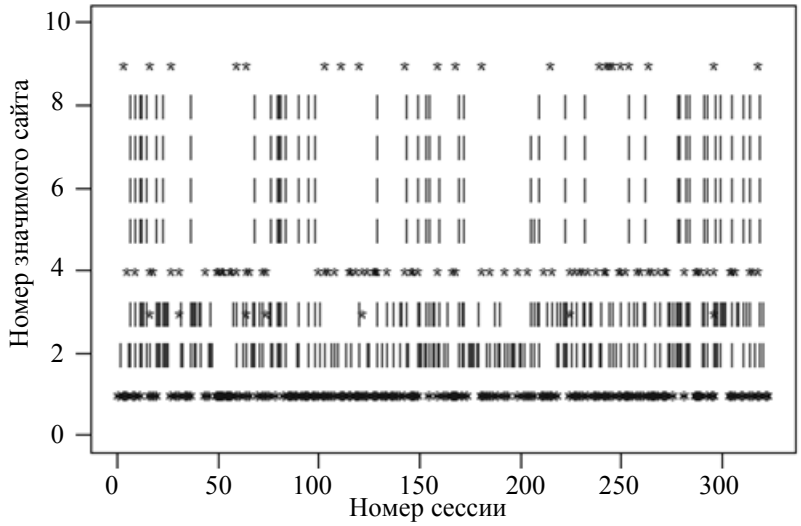

Рис. 1. Сессии пользователя: «*» – история посещений одного пользователя, «|» – другого

В качестве примера рассмотрим одну из «склеенных» cookies. На графике по оси абсцисс отметим номера сессий этой cookie, по оси ординат – номера ее значимых сайтов. Точку для сайта *i* сессии *j* отме-

чаем, если cookie посещала этот сайт в эту сессию. Звездочка или вертикальная черта обозначают, какому из изначальных «несклеенных» пользователей принадлежит данная сессия (рис. 1). Таким образом, все точки над одной сессией отмечены одинаковым символом. Из данного рисунка видно, что сайты 1, 4 и 9 посещал только пользователь, помеченный «*», сайты 2, 6, 7, 8 посещал только пользователь, помеченный «|», сайт 3 – в основном пользователь, помеченный «|».

Получившееся автоматическое разделение сайтов на кластеры имеет четко интерпретируемую трактовку, отвечающую требованиям решаемой задачи (рис. 2).

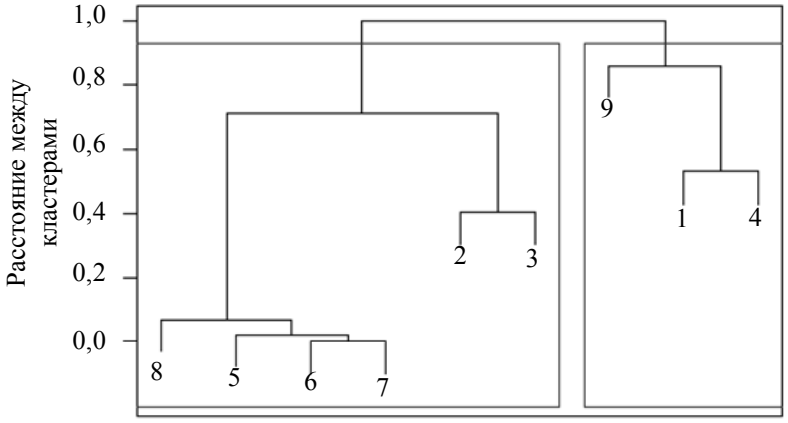

Рис. 2. Дендрограмма кластеризации «значимых» сайтов: 1–9 – номера сайтов

Результаты идентификации

Таким образом, для каждого cookie-идентификатора был сформирован список его значимых сайтов и произведена их кластеризация на две группы, которые обозначим как *A* и *B*. Для каждой его сессии было посчитано, сколько в ней сайтов из одного кластера и сколько из другого. Считаем, что сессия принадлежит пользователю U_A, если в ней встречается больше сайтов из кластера *A*, чем из *B*; сессия принадлежит пользователю U_B, если в ней встречается больше сайтов из кластера *B*, чем из *A*, остальные сессии считаем нейтральными. Почти все сессии получили метку U_A или U_B, нейтральные сессии далее были исключены из рассмотрения. В итоге все оставшиеся сессии имеют ровно по две метки – U_A или U_B и «*» или «|». Для каждой cookie была построена таблица сопряженности начальных и новых меток. В качестве количества правильно идентифицированных сессий для нее была взята большая из сумм чисел на главной и побочной диагоналях. Средний процент правильно идентифицированных сессий по всем пользователям в экспериментальных данных равен 76,7%.

Отметим, что процент правильно идентифицированных сессий пользователя зависит от их количества. Это видно из графика (рис. 3), где по оси абсцисс отмечено число сессий, а по оси ординат – средний процент правильно идентифицированных сессий для пользователя с их соответствующим количеством. Таким образом, для качественного разделения истории посещений сайтов многопользовательской cookie необходимо, чтобы у пользователя было по меньшей мере 100 сессий.

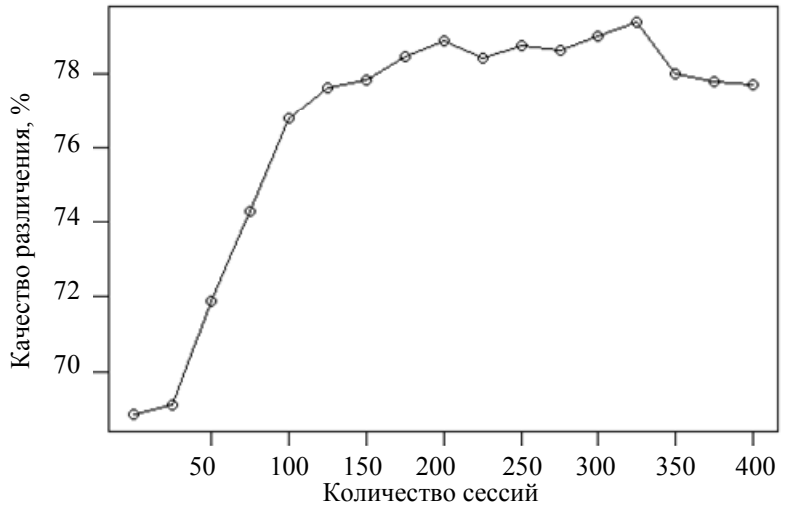

Рис. 3. Зависимость качества различения от количества сессий у пользователя

Заключение

Авторами разработан способ различения пары пользователей, осуществляющих доступ в сеть Интернет с одного компьютера и браузера, который основан на выявлении тех сайтов, которые пользователь посещает часто и регулярно. Для подтверждения работоспособности предложенного способа был подготовлен и проведен эксперимент с целью проверки качества различения более чем 260 000 пользователей на основе анализа их истории посещения сайтов. Процент правильно идентифицированных пользовательских сессий равен 76,7%, что является на сегодняшний день лучшим результатом. Дальнейшее исследование направлено на решение задачи определения факта, что компьютером и браузером пользуется несколько человек. Актуальной задачей является адаптация представленного способа для определения точного количества пользователей.

Литература

1. Understanding IP Addressing: Everything You Ever Wanted To Know [Электронный ресурс]. – Режим доступа: http://www.3com.com/other/pdfs/infra/corpinfo/en_US/501302.pdf, свободный. Яз. англ. (дата обращения 04.04.2013).
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, свобод-ный. Яз. англ. (дата обращения 04.04.2013).
3. Бессонова Е.Е., Зикратов И.А., Колесников Ю.Л., Росков В.Ю. Способ идентификации пользователя в сети Интернет // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – № 3 (79). – С. 133–137.
4. Бессонова Е.Е., Зикратов И.А., Росков В.Ю. Анализ способов идентификации пользователя в сети Интернет // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – № 6 (82). – С. 128–130.
5. Кантор И. Способы идентификации в интернете [Электронный ресурс]. – Режим доступа: http://javascript.ru/unsorted/id, свободный. Яз. рус. (дата обращения 04.04.2013).
6. Fulgoni G. When the Cookie Crumbles [Электронный ресурс]. – Режим доступа: http://www.comscore.com/Insights/Blog/When_the_Cookie_Crumbles, свободный. Яз. англ. (дата обращения 04.04.2013).
7. Lipkus A.H. A proof of the triangle inequality for the Tanimoto distance // Journal of Mathematical Chemistry. – 1999. – V. 26. – № 1–3. – P. 263–265.
8. Tan P., Steinbach M., Kumar V. Introduction to Data Mining. – Addison-Wesley, 2005. – P. 487–568.
9. Воронцов К. Лекции по алгоритмам кластеризации и многомерного шкалирования [Электронный ресурс]. – Режим доступа: http://www.ccas.ru/voron/download/Clustering.pdf, свободный. Яз. рус. (дата обращения 04.04.2013).
10. Jain A., Murty M., Flynn P. Data Clustering: A Review // ACM Computing Surveys. – 1999. – V. 31. – № 3. – P. 264 – 323.

Юрасов Дмитрий Сергеевич – Россия, Санкт-Петербург, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, аспирант, yurasov17@gmail.com

Зикратов Игорь Алексеевич – Россия, Санкт-Петербург, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, доктор технических наук, профессор, igzikratov@yandex.ru