УДК 004.931

АНАЛИЗ СПОСОБОВ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ В СЕТИ ИНТЕРНЕТ

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

Рассматриваются механизмы идентификации пользователей в сети Интернет. Предложен сравнительный анализ способов идентификации на основе регрессионного анализа и энтропийного подхода. Для проверки полученных результатов проведен вычислительный эксперимент.

Ключевые слова: идентификация, информативность, признак, кортеж, пользователь.

Введение

Одной из основных задач современной теории и техники автоматического управления является задача идентификации систем, т.е. определение структуры и параметров систем по наблюдениям [1].

В частности, в теории защиты информации актуальной является вопрос идентификации пользователя в сети Интернет. Актуальность данной темы обусловлена целесообразностью идентификации субъектов Сети при построении системы защиты информации для выявления нарушителей.

Целью данной работы является сравнительный анализ способов идентификации пользователей.

Для современных информационных систем применяются способы идентификации, основанные на хранении IP-адресов компьютеров посетителей и записи на компьютер пользователя данных Cookie. Однако оба способа не позволяют в ряде случаев достичь требуемой степени достоверности идентификации [2]. В работе [3] показан способ идентификации, основанный на регрессионном анализе, что позволяет осуществить рациональный выбор признаков, необходимых для повышения степени достоверности идентификации пользователя в сети Интернет. В процессе работы на основании сформированного признакового пространства с помощью регрессионного анализа были выбраны наиболее информативные признаки. Под информативностью признаков понимается степень влияния признака в кортеже идентифицируемого объекта на результат отождествления с имеющимися профилями пользователей.

Проведен сравнительный анализ метода идентификации пользователя по его кортежу путем сравнения его с накопленной базой. Результатом работы является кортеж, состоящий из восьми наиболее информативных идентификаторов: ETag, Supercookie, Cookie, MAC, IP, шрифты через Flash, плагины, шрифты через ActiveX. По сравнению с Cookie, такой кортеж обеспечивает в 6,3 раза большую информативность (4,35 против 0,69).

Кроме способа, основанного на методе регрессионного анализа, авторами исследован также метод формирования признакового пространства, основанный на вычислении количества энтропии [4].

Энтропия – это количество информации, приходящейся на одно элементарное сообщение источника, вырабатывающего статистически независимые сообщения. Для расчета энтропии Шеннон предложил уравнение

$$H(X) = -\sum_{i=0}^{N} P \log P(X_i),$$

где X – дискретная случайная величина с диапазоном изменчивости N; $P(X_i)$ – вероятность i-го уровня X [5]. Количество энтропии, которое содержит в себе признак, можно вычислить по формуле

$$\Delta S = \log_2 P_{x \in X}(x)$$

где $P(x)$ – вероятность появления значения x признака X.

Понятие энтропии может быть использовано для оценки признаков профиля пользователя. Значения энтропии отдельных признаков, которые приводит исследование, сделанное Electronic Frontier Foundation, отображены в таблице [6].

Для сравнительного анализа методов, основанных на регрессионном анализе и энтропийном подходе, был проведен эксперимент.

Наименование признака	Энтропия, бит
Заголовок Http User-Agent	10,0
Список установленных плагинов	15,4
Список установленных шрифтов	13,9
Поддержка supercookies	2,12
Заголовок Http Accept	6,09
Временная зона	3,04
Включенность cookies	0,353

Таблица. Количество энтропии информативных признаков (по данным Electronic Frontier Foundation)

Проведение эксперимента

Авторами был сделан сравнительный анализ двух кортежей, полученных в указанных выше работах, по степени их достоверности и скорости идентификации пользователей. Для этого был проведен эксперимент с целью вычисления степени достоверности идентификации и времени работы. В качестве результатов эксперимента получены временны́е характеристики работы двух кортежей, а также зависимость количества идентифицированных пользователей от уровня шума для обоих кортежей.

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.

Целью эксперимента являлось определение степени достоверности работы двух кортежей и их скорости обработки данных. Результаты эксперимента представлены на графиках (рис. 1).

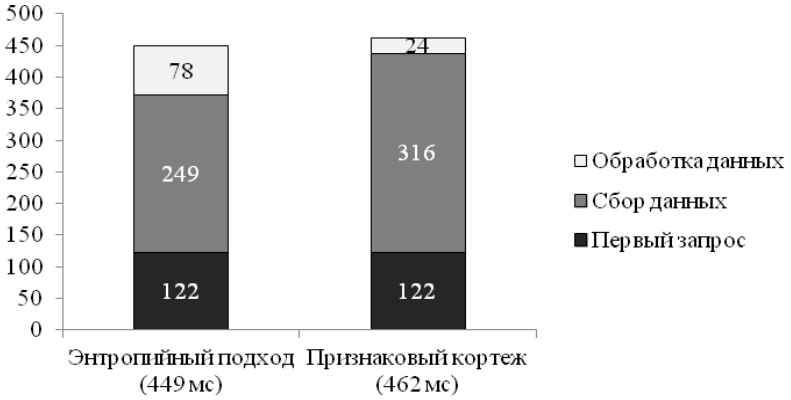

Рис. 1. Сравнение временных характеристик процессов обработки кортежей

В представленных на графике результатах можно выделить три временных интервала:

1. первый запрос – сколько времени занимает загрузка страницы, к которой подключен скрипт с данными;
2. сбор данных – время от начала работы скрипта до момента отправки их на сервер;
3. обработка данных – время с момента отправки данных на сервер до получения результата идентификации.

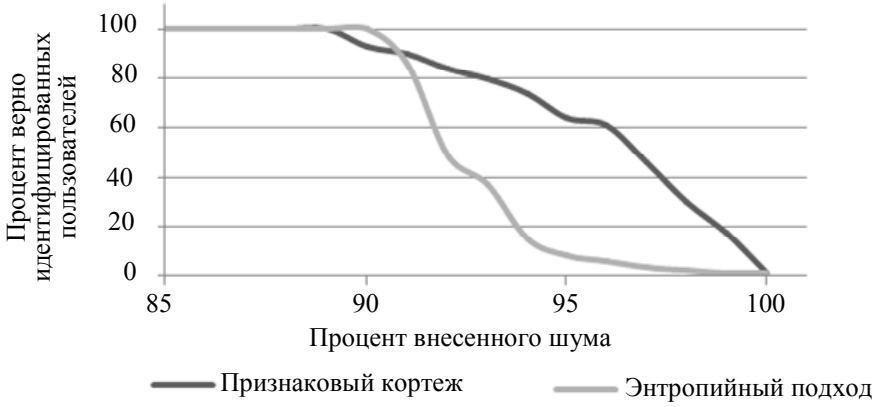

Рис. 2. Сравнение достоверности идентификации для различных способов

Данные получены при усреднении 10 000 запросов. Очевидно, что длительность первого запроса к странице во всех случаях одинакова. Так как у энтропийного кортежа меньше признаков, сбор данных происходит быстрее. Подсчет энтропии занимает больше времени, чем предложенный авторами метод идентификации. Таким образом, экспериментально полученная скорость работы обоих методов отличается незначительно. При использовании признакового кортежа время увеличивается на 2,9%.

Эксперимент показал, что при внесенном шуме, составляющем 89%, оба подхода демонстрируют одинаковую эффективность. Однако при внесенном шуме более 90% энтропийный подход резко ухудшает свои результаты и при 95% шума может идентифицировать менее 10% пользователей, тогда как метод идентификации, предложенный в работе [3], позволяет идентифицировать в 6 раз больше пользователей (рис. 2).

Заключение

Таким образом, для агрессивной среды более подходящим для использования является предложенный авторами метод, так как он показал более высокую степень достоверности. Для неагрессивной среды представляется возможным использовать энтропийный подход, так как он имеет более высокое быстродействие. Однако следует учитывать, что разность в быстродействии процессов идентификации при реализации рассматриваемых методов незначительна.

Литература

1. Цыпкин Я.З. Информационная теория идентификации. – М.: Наука. ФИЗМАТЛИТ,1995. – 336 с.
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. 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. Eckersley P. How Unique Is Your Web Browser? – Electronic Frontier Foundation, 2010. – 19 p.
5. Шеннон К. Работы по теории информации и кибернетики. – М.: Издательство иностранной литературы, 1963. – 830 с.
6. Eckersley P. A Primer on Information Theory and Privacy. – Electronic Frontier Foundation. – 2010. – 25 p.

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

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